

UNIVERSITY OF BAHRAIN

College of Information Technology



Cancer Prediction based on DNA Next Generation Genomic Sequencing using Machine Learning Algorithms

A Thesis Submitted in Partial Fulfillment of the Requirements for the
M.Sc. Degree in Information Technology

Submitted by
Marwa Assim Ayfan Lateef
20190175

Supervised by
Dr. Riadh Bin Mohammad Ksantini
Associate Professor

Dr. Fatema Abdulqader Yousif Albalooshi
Assistant Professor
University of Bahrain

Kingdom of Bahrain
March 2021



Dissertation Deposit Form

Author:- **Marwa Assim Ayfan Lateef**

Thesis Title:- **Cancer Prediction based on DNA Next Generation Genomic Sequencing using Machine Learning Algorithms**

Year:- 2021

College:- **Information Technology** Email:- **20190175@stu.uob.edu.bh**

Department:- **Computer Science** Contact No:- **33302902**

Author's Declaration: I agree to the following conditions:

The above thesis will be available (in the University of Bahrain library and externally) and reproduced as necessary at the discretion of the University of Bahrain. It may also be digitized by the University of Bahrain and made available via the university's repository or via other parties on the Internet.

This condition shall apply to ALL copies including electronic copies.

The above thesis has been provided on the understanding that it is copyright-protected material and that no quotation from this thesis may be published without proper acknowledgment.

Signature of the Author:-

A handwritten signature in black ink, appearing to read "Marwa Assim Ayfan Lateef".

Date:- **03/11/2021**

Declaration

I declare that this work is the result of my own investigation and that it has not already been accepted in substance for any degree, nor is it currently submitted for any degree.

Signed: **Marwa Assim Ayfan Lateef** *Marwa Assim* Date: 01/11/2021
(Candidate)

The Thesis defense committee considers the thesis titled:

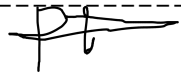
Cancer Prediction based on DNA Next Generation Genomic Sequencing using Machine Learning Algorithms.

Submitted by Marwa Assim Ayfan to the College of Information Technology is satisfactory and acceptable for the Master of Science in Information Technology.

The Defense Committee:

Dr. Riadh Bin Mohammad Ksantini

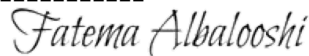
(Supervisor)



Department of Computer Science
University of Bahrain

Dr. Fatema Abdulqader Albalooshi

(Co-Supervisor)



Department of Computer Engineering
University of Bahrain


Dr. Mohammed Ghazal

University of Abu Dhabi

(Member)

Dr. Alauddin Yousif Alomary

(Member)



Department of Computer Engineering
University of Bahrain

Abstract

Cancer is a dangerous disease that threatens human life and leads to severe side effects and consequences, and in many cases, it leads to death. Cancer occurs when the cells start to divide and grow exponentially and uncontrollably. The problem is that cancer is often not discovered early in its first stage until after a long period has passed. Hence, the treatment, in this case, is more complicated and sometimes is not possible. Different diseases, including cancer, progress because of the mutations in the Deoxyribonucleic Acid (DNA) genomic sequence. Genes' mutations Next-Generation Sequencing (NGS) technology advancement enabled scientists and bioinformaticians to study genomic sequencing, which helps identify the effects of gene mutations on developing different types of diseases. This thesis focuses on predicting and classifying different cancers based on NGS technology by utilizing Ribonucleic Acid Sequence (RNA-Seq) gene expression data using various machine learning algorithms. This research aims to find the most potential biomarkers for various cancers that significantly help in personalized medicine. Four different approaches and models have been proposed in this study, namely Mutual Information-Support Vector Machine, Recursive Feature Elimination-Support Vector Machine, Random Forest. The fourth model is a novel ensemble approach that discovers the most relevant genes causing a specific type of cancer. Moreover, four datasets (COAD, LUAD, BRCA, and Combined dataset from the first three datasets) were applied and tested with the developed models, and the results were justified and validated. A comparison study of the obtained results with the existing research has been done. The experimental results within accuracy scores of up to 99.4% show that the proposed models compared to the state-of-the-art methods, efficiently predict and classify cancer and discover the most potential genes to develop a specific type of cancer.

*To my mother, daughter, and the
spirit of Dr. Khalaf, may Allah have mercy on him.*

Acknowledgments

Foremost, I would like to express my special thanks to Allah, who blessed me with health and the ability to complete my thesis successfully. First, I would like to express my sincere gratitude to my supervisor, Dr. Riadh Ksantini, for accepting being my supervisor, trusting my thesis idea, providing assistance and support, always being present when needed, and his continuous encouragement. I would also like to thank Dr. Fatema Al Balooshi, my co-supervisor, who helped me and encouraged me a lot, and trusted me. Secondly, I would like to thank Dr. Alaa for his continuous support, efforts, and guidance through the master's study. My sincere thanks go to Dr. Mustafa Al Hammadi for his continuous support, assistance, encouragement, and motivation to do research and self-development. Besides, thanks to Dr. Hessa Aljunaid for her support in the research field. Thanks, Dr. Jaffla Al Ammari, who was embracing us as a mother and motivating us. I would also like to thank Dr. Qasim Obiedat, Dr. Abdullah Qaddomi, Dr. Mohsen Roshaid, and Dr. Maan, who supported me during the master's journey. Great thanks and gratitude to Dr. Wael for his continuous support and effort in managing the thesis process. Great thanks to the University of Bahrain for accepting me to complete my master's degree and providing all the required services and facilities. Finally, I would like to give my heartfelt thanks to all my family and friends who supported me and encouraged me throughout my thesis. Without their support, I would not have been able to complete my master's degree. "The results shown here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>."

Table of contents

Defense committee’s approval.....	A
Abstract	C
Dedication	D
Acknowledgments	E
Table of contents	F
List of Tables	H
List of Figures	I
List of Abbreviations	J
List of Appendices	L
Chapter 1: Introduction	2
1.1 Overview	2
1.2 Background of the Study.....	2
1.3 Problem Definition.....	3
1.4 Research Significance	4
1.5 Research Objectives	4
1.6 Methodology	5
1.7 Contribution	6
1.8 Research Scope and Limitations	7
1.9 Thesis Outline	8
Chapter 2: Literature Review	11
2.1. Overview	11
2.2. Microarray Profiling vs. Next-Generation Sequencing (NGS)	11

2.2.1.	DNA Microarray Profiling.....	11
2.2.2.	Next-Generation Sequencing (NGS)	13
2.3.	Challenges in Cancer prediction research using Machine Learning.....	19
2.3.1.	High Dimensionality	19
2.3.2.	Class Imbalance	25
2.4.	Summary	28
Chapter 3: The Proposed Approach.....		31
3.1.	Overview	31
3.2.	Cancer Prediction and Early Detection	31
3.2.1.	Problem Description	31
3.2.2.	The Proposed Models	36
3.3.	Summary	55
Chapter 4: Experimental Results.....		57
4.1.	Overview	57
4.2.	Datasets	57
4.3.	The Results Analysis and Discussion	61
4.3.1.	Implementation Parameters	61
4.3.2.	Evaluation Criteria	62
4.3.3.	Experimental Results without Feature Scaling	63
4.3.4.	Experiment Results with Feature Scaling	64
4.3.5.	Comparative Study.....	75
4.3.6.	Biological Analysis of the Potential Biomarkers for Cancers	78
4.4.	Summary	82
Chapter 5: Conclusion and Recommendations		84
5.1.	Overview	84
5.2.	Conclusion.....	84
5.3.	Recommendation.....	86
Appendices.....		106

List of Tables

Table (3.1): List of imported libraries	38
Table (3.2): List of MI implementation Parameters	42
Table (3.3): List of RFE implementation Parameters.....	47
Table (3.4): List of RF implementation Parameters	52
Table (4.2): List of MI-SVM model implementation Parameters Values.....	61
Table (4.3): List of RFE-SVM model implementation Parameters Values	62
Table (4.4): List of RF model implementation Parameters Values	62
Table (4.5): Accuracy results of COAD Dataset without Feature Scaling.....	64
Table (4.6): MI-SVM Model Classification Results Report	67
Table (4.7): RFE-SVM Model Classification Results Report.....	70
Table (4.8): RF Model Classification Results Report	72
Table (4.9): Accuracy results using top selected genes.....	73
Table (4.10): Number of top selected genes	74
Table (4.11): Accuracy results using all the genes–No Feature Selection.....	74
Table (4.12): Implementation Processing Time of all the Models in Seconds	75
Table (4.13): Comparison of the Proposed Models with other Studies	77

List of Figures

Figure (3.1): Central Dogma system of Molecular Biology.....	33
Figure (3.2): General workflow of Machine Learning Model.....	36
Figure (3.3): The workflow of the proposed Machine Learning Models.....	37
Figure (4.1): Accuracy of MI-SVM model in COAD Dataset.....	65
Figure (4.2): Accuracy of MI-SVM model in LUAD Dataset	65
Figure (4.3): Accuracy of MI-SVM in BRCA Dataset	65
Figure (4.4): Accuracy of MI-SVM model in Combined Dataset.....	65
Figure (4.5): Accuracy of RFE-SVM in COAD Dataset.....	68
Figure (4.6): Accuracy of RFE-SVM in LUAD Dataset.....	68
Figure (4.7): Accuracy of RFE-SVM model in BRCA Dataset.....	68
Figure (4.8): Accuracy of RFE-SVM model in Combined Dataset	68
Figure (4.9): Accuracy of RF model in COAD Dataset	71
Figure (4.10): Accuracy of RF model in LUAD Dataset	71
Figure (4.11): Accuracy of RF model in BRCA Dataset	71
Figure (4.12): Accuracy of RF model in Combined Dataset.....	71
Figure (4.13): Venn Diagram of COAD Colon Cancer most Common Genes.....	80
Figure (4.14): Venn Diagram of LUAD Lung Cancer most Common Genes	80
Figure (4.15): Venn Diagram of BRCA Breast Cancer most Common Genes.....	81
Figure (4.16): Venn Diagram of Combined Cancers Most Common Genes	81

List of Abbreviations

AE	Auto Encoder
AI	Artificial Intelligence
CSV	Comma Separated Value
CUPS	Cancer of Unknown Primary Site
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
DT	Decision Tree
ENC	Empirical Negative Control
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
LR	Linear Regression
LinearSVC	Linear Support Vector Classifier
MI	Mutual Information
mRMD	maximum Relevance Minimum Distance
NB	Naïve Bayes
NGS	Next-Generation Sequencing
NN	Neural Network
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RF	Random Forest
RNA	Ribonucleic Acid
SPL	Self Paced Learning
SVC	Support Vector Classifier
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TN	True Negative

TP	True Positive
RFE	Recursive Feature Elimination
RUV	Remove Unwanted Variation

List of Appendices

Appendix (1)	106
Appendix (2)	111
Appendix (3)	114

Chapter 1

Introduction

- 1.1. Overview
- 1.2. Background of the Study
- 1.3. Problem Definition
- 1.4. Research Significance
- 1.5. Research Objectives
- 1.6. Methodology
- 1.7. Contribution
- 1.8. Research Scope and limitations
- 1.9. Thesis Outline

Chapter 1: Introduction

1.1 Overview

Cancer is one of the most common diseases that lead to death or harmful consequences for the patient, for which no definitive treatment has been available so far. From this standpoint, the need for an effective solution to detect the tumor's site of origin and cancer type in an early stage is essential. Since metastasis tumors, where the tumor spread from its primary site of origin to different parts of the body, are challenging to treat. Early cancer detection could distinguish tumors at early stages when results are guaranteed and treatment is less painful. DNA genomic sequencing and profiling help predict the probability of different types of cancers by searching for specific genetic changes, chromosomes, and proteins. These changes are called mutations. In this study, various machine learning techniques are proposed to help predict cancers by analyzing the genomic structure of the DNA and, more specifically, the RNA-Seq gene expression. Moreover, this study focuses on discovering the most potential biomarkers (genes) causing a specific cancer type that significantly helps with personalized medicine.

1.2 Background of the Study

Genomic sequencing has become a popular and exciting research field in recent years. Utilizing genome sequencing in cancers diagnosing and treatment is developing rapidly. The advancement of Next-Generation Sequencing (NGS) techniques (Mardis & Wilson, 2009) that provide high sequencing quality of DNA and RNA genomics encourages more research in cancer. Comparing the NGS sequencing techniques with the old technology in terms of quality, technology involved, and cost required for

sequencing one human genome increases the possibility of achieving more discoveries and advancements in the field of bioinformatics. The old sequencing technologies like targeted gene sequencing using Polymerase Chain Reaction (PCR) and Sanger Sequencing (Michal Janitz, 2011) are extensively expensive, where one human genome sequencing cost around \$1.5 million. In contrast, NGS costs around less than \$1000 with the Illumina platform that takes up to 2.5 terabytes of space (Brendan P. Hodkinson Elizabeth A. Grice, 2015). Remarkable advances occurred during the last years in studying the genetic mutations in the DNA and their effect on the human health system (Cristiano et al., 2019).

This research is based on NGS technology in which RNA-Seq gene expression data is used to implement the proposed machine learning algorithms for predicting different types of cancers and finding the most potential biomarkers for cancer.

1.3 Problem Definition

The problem considered in this study is cancer disease and its threat to human life. Early detection of cancer decreases the harmful effects and increases survival opportunities. Due to cancer's nature, in which it cannot be detected at early stages. The cancerous cells are hiding around the normal cells, keeping them undetectable. This study proposes different machine learning techniques that efficiently predict and classify different types of cancers using NGS sequencing techniques. Besides implementing a novel ensemble approach in finding the most potential biomarkers for cancer, that can help in personalized medicine. This study mainly focuses on considering the factors and challenges that affect cancer prediction accuracy, which was not considered in the previous studies. Among these problems is feature scaling of the DNA sequencing data, where the

selection of feature scaling technique affects gene expression correlation as was done by (Sherafatian & Arjmand, 2019). The selection and the implementation of feature selection techniques and the classification algorithm may degrade the prediction accuracy, as was done by (Zhao et al., 2019) and (Cueto-lópez et al., 2019).

1.4 Research Significance

Cancer is a harmful and dangerous disease that causes severe problems for the patients or may lead to death in many cases, especially if left undetected in its early stages. NGS for DNA and RNA-Seq gene expression is a new technology that can help in many bioinformatics fields. Recently, many studies conducted adopting the genomic sequencing technique in solving different medical problems by detecting particular changes or mutations that causes a specific disease. In this research, we focus on improving the accuracy results in predicting a primary tumor for a specific type of cancer. This aim can be achieved using RNA-Seq gene expression technology by utilizing machine learning techniques. Moreover, the proposed models can detect the most potential biomarkers in a particular cancer type that significantly helps with personalized medicine. The physicians are using biomarkers tests in identifying the gene mutations of each patient, which helps in the follow-up and treatment of the patients individually (Goossens, Nakagawa, Sun, & Hoshida, 2015).

1.5 Research Objectives

The thesis's primary goal is to optimize the results of an existing model or propose a novel model to predict cancers using DNA genomic sequencing and profiling. Based on the literature review, we found a lack of tackling two main limitations in this research field. Due to the massive

size of the whole genomic sequencing dataset, different challenges face the researchers while analyzing the dataset and training the model like high dimensionality and class imbalance. Thus, in this thesis, we focus on solving data high dimensionality challenges by proposing different machine learning techniques for classifying and predicting different types of cancers efficiently.

The following are the various objectives of this thesis work: -

1. To survey the literature for similar work.
2. To develop different models for predicting and classifying different types of cancers based on RNA-Seq gene expression using Machine Learning algorithms. The proposed models' results to be compared with existing models in terms of specificity, sensitivity, precision, and F1 measure and consider a novel model based on RNA-Seq gene expression analysis for cancer prediction.
3. To find the most potential biomarkers (genes) that are highly contributing to a specific type of cancer by reducing high data dimensionality in the RNA-Seq gene expression. In this study, biomarkers and gene terms are used interchangeably to express the genes in the DNA that are highly correlated to the development of cancers.

1.6 Methodology

The research method that is used to conduct the thesis is a quantitative approach. Following are the various phases of the designed methodology:

1. Literature concerning DNA and RNA genomic sequencing and profiling in predicting different types of cancers has been reviewed to conduct the thesis research.
2. Feature Selection: More efficient algorithms capable of predicting cancer have been proposed, like different feature selection techniques that aim to reduce the high dimensionality related to the RNA-Seq gene expressions.
3. Classification: In this step, the classification models are used to predict and classify cancer. Different techniques of machine learning algorithms have been implemented to classify different types of cancer efficiently. Besides, an ensemble model has been proposed to discover the most potential biomarkers for the development of different types of cancers.
4. Performance Analysis: In this phase, the proposed models' performance is analyzed and evaluated in terms of sensitivity, specificity, precision, and F1 measure.
5. Biological analysis on the most common potential biomarkers (genes) has been done. The analysis purpose is to confirm further the proposed models' efficiency in detecting the most correlated genes to cancer development.

1.7 Contribution

This dissertation illustrates that the key to efficient classification and diagnosis of cancer using machine learning is by utilizing NGS technology and different machine learning techniques.

The principal contributions of this thesis study are:

1. Various models have been proposed and implemented to predict and diagnose cancer efficiently compared to state-of-the-art studies. The models are implemented by utilizing the new genomics sequencing technology using RNA-Seq gene expression. In this study, three datasets are utilized (COAD, LUAD, BRCA, and fourth dataset that combine the first three datasets to form a multi-cancer dataset).
2. A list of most contributing genes to a specific cancer type is obtained from each model's output.
3. A novel approach is proposed, which combines the results of the various proposed and implemented models to find the most potential biomarkers (genes) that are highly correlated to developing cancer. This objective can be achieved by reducing the high dimensionality in the massive RNA-Seq gene expressions.
4. These biomarkers significantly help in personalized medicine. Personalized medicine can be defined as if two patients have the same cancer type; they can be treated differently. The physician prescribes different medicine for each patient according to the mutations and gene expression levels in their genomic sequence.
5. Contribution to the field of bioinformatics by computationally discovering cancer-causing genes whose discovery in the laboratory requires time, effort, and high cost to study each gene and its effect on a specific cancer type. Determining the biomarkers for each cancer type helps in diagnosing cancer from the genome sequence of persons.

1.8 Research Scope and Limitations

Recently many research studies are conducted that adopting machine learning algorithms in the bioinformatics field. According to the

advancement in genomic sequencing techniques, studying the genetic effect on different health issues increases. These studies involve using different machine learning algorithms to predict specific disease probability and diagnose the availability of different diseases according to the mutations in DNA genes sequence or the RNA expression levels. Due to the high dimensionality involving genomic sequencing data, this thesis study focuses on solving this challenge. Different machine learning algorithms and techniques were adopted and implemented. The results are promising compared with the results obtained from the state-of-the-art machine and deep learning techniques.

The research scope includes the following areas:

1. Present the state-of-the-art cancer prediction methods based on genomic sequencing using machine learning algorithms from the literature.
2. Predict and classify specific primary tumors from solid normal tissue using RNA-Seq gene expression data by proposing different machine learning techniques.
3. Find the most potential biomarkers for specific types of cancers.
4. Compare the results of the proposed models with the results of recently conducted research.

1.9 Thesis Outline

The thesis report structure and design are illustrated as follows:

- Chapter 1: This chapter includes an introduction and a background to the thesis study. It includes an overview, background, problem

definition, research significance, objective, methodology, contribution, and research scope and limitations.

- Chapter 2: This chapter introduces the Literature Review that highlights the recent relevant studies of the research. It contains different approaches proposed by the researchers for predicting cancer using different sequencing techniques and machine learning algorithms.
- Chapter 3: This chapter highlights the methodology used for the implementation of the developed models. It includes the proposed approaches and the workflow of their implementation.
- Chapter 4: This chapter analyses the research results and findings. It explains and discusses the validation of the obtained results and comparing them with other research in the same area.
- Chapter 5: This chapter concludes the research and provides a recommendation study in the same field. It summarizes the main finding and future research directions.

Chapter 2

Literature Review

- 2.1. Overview
- 2.2. DNA Microarray Profiling vs. Next-Generation Sequencing (NGS)
- 2.3. Challenges in Cancer prediction research using Machine Learning
- 2.4. Summary

Chapter 2: Literature Review

2.1. Overview

Cancer is a complex and challenging disease that causes severe damage and consequences to the patient. Diagnosing cancer is a tedious, complicated, and time-consuming task where the pathologist examines the patient's biopsy under a microscope to categorize the tumor type. Genomic mutations and alterations are correlated to a specific tumor type. Hence, in this literature, different studies and techniques are presented to build a basic knowledge about the potentials in predicting cancer types, using DNA and RNA-Seq gene expression techniques. The main techniques are based on two main categories, namely Microarray profiling and NGS.

2.2. Microarray Profiling vs. Next-Generation Sequencing (NGS)

2.2.1. DNA Microarray Profiling

Microarray is a set of microscopic DNA or DNA chips tied to a robust surface. Microarray profiling is widely used to extract or genotype the genome expression simultaneously. However, it focuses on targeted sequence rather than the whole DNA genome or RNA-Seq gene expression. It is used to determine whether a patient's DNA contains gene mutation for detecting specific types of diseases according to these mutations (Sayed, Nassef, Badr, & Farag, 2019). Several studies have previously explored this technique in predicting different types of cancers by identifying specific gene mutations. (Statnikov, Wang, & Aliferis, 2008) conducted a comprehensive comparison in applying machine learning algorithms for predicting cancer type. The study aimed to evaluate

the performance of predicting different gene mutations for specific types of cancer using the DNA Microarray gene expression dataset. Two machine learning algorithms were applied in their study, namely Support Vector Machine (SVM) (Evgeniou & Pontil, 2001) and Random Forest (RF) (Fawagreh, Medhat Gaber, Elyan, & Gaber, 2014). The results revealed that the SVM classifier performed better than RF in predicting cancer type according to the gene mutations in twenty-two DNA Microarray datasets. The results were tested using two approaches. The first one is building the model with no gene selection techniques where no data preprocessing was performed on the dataset. The second approach is selecting the best performing genes in predicting a specific type of cancer. In both approaches, SVM performed better than the RF classifier. However, more efficient gene selection techniques should be adopted and evaluated to measure the classifiers' accuracy.

Moreover, (Nannini et al., 2009) conducted a comprehensive review of state-of-the-art techniques adopted DNA Microarray to predict colorectal cancer. The review study focused on implementing Microarray gene expression data to discover gene mutations and patterns. The study is related to diagnosing Colorectal cancer and classifying important molecular factors, helping the pathologist in their treatment decisions. The study emphasized that recent studies adopted NGS techniques to analyze DNA genomics. It focuses on the whole DNA genome rather than pre-specified and targeted genes to correctly predict the most predictive genes.

Similarly, (Colombo, Milanezi, Weigelt, & Reis-Filho, 2011) conducted a review study to analyze the importance of adopting DNA Microarray technology in predicting breast cancer. The fact that cancer has distinctive histopathological and molecular features and signatures led to discovering

multi-genes patterns in predicting breast cancer and response to its therapy approach.

Although studies showed that applying DNA Microarray sequencing technology helps significantly predict different types of cancers efficiently (Colombo et al., 2011), different studies demonstrated that applying a small number of samples of the Microarray dataset on the classification model affects the prediction accuracy results. Hence, vast samples should be included in the training and validation datasets. Thus, genome mutations and patterns can be efficiently identified, and the misclassification issue can be avoided (Michiels, Koscielny, & Hill, 2005).

2.2.2. Next-Generation Sequencing (NGS)

Next-Generation Sequencing (NGS) is the new technology of sequencing a whole DNA human genome or RNA gene expression by fragmenting them. Millions of small fragments are then sequenced in parallel. Resulting in a whole-genome DNA or RNA sequence that can be applied in different fields and applications by researchers in computer science and bioinformatics field. Recently, various studies started exploring the NGS technique in predicting different types of cancers by identifying the site of origin and classifying the most predictive gene mutations related to a specific cancer type.

A recent study (Hooshmand, 2020) proposed a novel model for predicting different types of tumors using a machine learning technique. The proposed model aimed to evaluate the application of gene transcriptomic data using the DNA-NGS sequencing technique in predicting cancer types. In this model, Whole Exome Sequencing of RNA of twenty-two types of cancers extracted from The Cancer Genome Atlas (TCGA) project and samples of healthy tissues collected from the Genotype-Tissue Expression (GTEx)

project. Different samples from different people were also collected for testing and evaluating the developed model. The model was also built to predict the new patients' status and the death estimation because of the cancer disease. Thus, the application of NGS provides a promising approach in the field of oncology. Naïve Bayes (NB) (Lee et al., 2001), SVM, Decision Tree (DT), RF, Linear Regression (LR) (Schneider, Hommel, & Blettner, 2010), K- Nearest Neighbors (KNN) (Cunningham & Delany, 2007), and K-means clustering (Bock, 2007) machine learning algorithms were applied in this model. The results obtained that the model is efficiently predicting if the patient is healthy or has cancerous cells. However, the model was not tested on patients' organs. So it should be tested on different datasets to find the most relevant biomarkers in the patient's blood.

A different study done by (N. Wan et al., 2018) evaluated the application of Cell-Free DNA (cfDNA), blood-based DNA samples using machine learning algorithms to predict colorectal cancer. The study aimed to evaluate the possibility and accuracy of the prediction of whole-genome sequenced cfDNA. Pre-processing techniques were applied to the proposed model first. Principal Component Analysis (PCA) (Pothuganti, 2020) was used to eliminate high-dimensional data and select the most relevant gene mutations. LR and SVM classifiers were applied to classify the cancer type. The obtained results show relatively high accuracy results. However, the model needs to be trained and evaluated by implementing different feature selection techniques to select the most relative genes efficiently.

Finally (Jiao et al., 2020) developed a model for evaluating whether whole DNA genome sequencing can be used with machine learning techniques to determine the site of origin and histology of tumors using somatic mutation patterns. The model was trained using individual features of the dataset and

a combination of features to evaluate the importance of specific features in the prediction accuracy. In this study, a random forest classifier and deep neural networks were used to evaluate the accuracy of predicting cancer type. Deep neural networks achieved higher accuracy scores than random forest classifiers.

Histopathologic tissue examination is still the primary method for diagnosing cancer types (Aswathy & Jagannath, 2017). A biopsy sample is extracted from the tissue and examined under a pathologist's microscope when a patient notices signs and symptoms. If the cells are diagnosed as cancerous, the cancer sub-type will be categorized according to the site of origin. In most cases, the patients have cancer of unknown primary, where the site of origin cannot be determined by examining metastatic cancer tissues. Most of the patients with cancer of unknown primary have little chance of survival (Çelik et al., 2018).

Numerous studies have been conducted to address the cancer type and site of origin identification. With the rapid development of Artificial Intelligence technologies and the implementation of various machine and deep learning algorithms in different applications, cancer prediction has progressed increasingly in recent years. (Penson et al., 2020) developed a model to predict and diagnose cancer types between different types of cancers in an advanced stage. The study aimed to evaluate the model in predicting the tissue of origin for different types of cancers. RF machine-learning classifier has been applied to classify and predict cancer type. The classifier's scores then calibrated using Multinomial Logistic Regression (MLR) (Aroef, Yuda, Rustam, & Pandelaki, 2019). The prediction accuracy of the developed model reached 73.8%. Another study was done by (Soh, Szczurek, Sakoparnig, & Beerenwinkel, 2017) to evaluate the potential of predicting the cancer type from a small set of gene alterations

of the tumor DNA cells like somatic point mutations and copy number alterations. This study focused on evaluating the use of as few genetic features as possible in predicting the cancer type using machine learning classifiers. This study aimed to develop a framework for identifying cancer's primary site for the patients diagnosed with cancer, but the site of origin remains unknown. Three machine learning algorithms were applied in this study, including RF, LR, and SVM. SVM classifier yielded the highest accuracy score compared with RF and LR.

On the other hand, (Kim, Rachid Zaim, & Subbian, 2020) conducted a statistical assessment and analysis of the state-of-the-art machine learning algorithms. The study aimed to evaluate the application of machine learning techniques from the perspective of interpretability and reproducibility for detecting the existence of different types of cancers. Different machine learning algorithms were applied, like RF and Gradient Boosting (GB) (Natekin & Knoll, 2013) machines. This study assumed that there is no one specific model that the clinical analysis can depend on in the analysis and diagnosis of cancer. Accordingly, it suggests that more transparent studies should be conducted to improve the reliability in applying machine learning techniques in biomedical analysis.

Therefore, several studies integrated different models to increase the accuracy scores. For instance (Abraham et al., 2021) proposed a novel approach using machine learning algorithms like RF, SVM, LR, KNN, NB, Quadratic Discriminant Analysis (QDA) (Tharwat, 2016), and Gaussian Processes (GP) (Rasmussen & Kuss, 2004) to diagnose tissue of origin of different types of cancers. The results show that the model obtained high accuracy rates in detecting twenty-one types of cancers. However, a

limitation to this model is that it is not trained on specific tumors that should be evaluated in a traditional pathologic test.

Similarly, (Ayyad, Saleh, & Labib, 2019) proposed cancer classification model by implementing a modified version of the KNN algorithm (MKNN). The model was built by implementing two steps. The first step is Smallest MKNN, and the second step is Largest MKNN. A new weighting process has been applied to extract healthy neighbors in the dataset. The model performed better in both Smallest MKNN and Largest MKNN scenarios than the KNN classifier. However, the model should be tested using real datasets to reveal the real model efficiency in predicting the cancer type.

None of the machine learning classifiers is entirely accurate in predicting the cancer type from the studies mentioned above. Besides, considering the massive size of the datasets and the cancer prediction process's complexity, a more efficient and accurate technique is needed. Unlike most machine learning classifiers that only address the linear relationships between the classifiers and need manual intervention, deep learning methods can efficiently learn the non-linear relationships between the classifiers and automatically detect the most predictive values from the large dataset. Various studies applied deep learning techniques for predicting cancer type and site of origin.

(Ching, Zhu, & Garmire, 2018) proposed a deep-learning-based approach to predict the survival probability. Cox Regression (Chen et al., 2019) and Neural Networks (NN) (Kucharska, 2019) models were adopted. Cox Regression model used as the output layer of the NN to build the model. One hidden layer was used for training the NN classifier to reduce its

complexity. The data drop-out technique was also implemented to avoid overfitting by dropping the data with less prediction value. (Katzman, Shaham, Cloninger, Bates, & Jiang, 2018) also utilized the NN capabilities by developing a Deep Learning model based on NN and implementing the same approach followed by (Ching et al., 2018). The forward NN model was built by tuning some common hyperparameters of the NN classifier to predict the patient's survival probability. The results show that applying NN in predicting cancer survival achieved high accuracy results. However, in this approach, one NN structure was implemented to test the prediction's accuracy level. More NN should be considered to test the model performance in terms of cancer prediction.

On top of that, different studies (Jing et al., 2019) proposed the same approach by implementing NN but using more complex Forward Deep NN (DNN). Three to four hidden layers were added to train the model and increase the prediction accuracy. In this model, loss function and ranking loss techniques were implemented for optimizing the Forward NN classification. The results showed better performance than the previous two NN models (Ching et al., 2018) and (Katzman et al., 2018).

Another study done by (Xiao, Wu, Lin, & Zhao, 2018a) proposed a deep learning ensemble multi-model approach for predicting cancer type. The study aimed to increase the accuracy level of the prediction by integrating several machine learning algorithms and ensemble the classifiers' results as an input to the DNN classifier. Five machine learning classifiers applied KNN, SVM, DT, RF, and Gradient Boosting Decision Tree (GBDT). The output of these classifiers then ensembled as input to the DNN algorithm to obtain higher accuracy results. The results show that utilizing the ensembled multi-model technique improved the accuracy score for cancer type prediction.

A different study was done by (Xiao, Wu, Lin, & Zhao, 2018b) proposed Stacked Sparse Auto Encoder (SSAE) technique for predicting breast cancer, where Auto Encoder is used for feature selection. The result of the proposed model was 96.23%. Deep learning techniques are characterized by their high ability to determine the related features. Still, it requires a high processing time and a large number of samples.

Since cancer is a genetic disease (Wishart, 2015), the dataset's size related to cancer prediction is enormous due to the size of the DNA genome sequence. It contains a massive number of somatic genetic mutations. Many cancer prediction research challenges using machine learning techniques are presented, like high dimensionality, imbalanced class. Most of the studies mentioned in the literature did not consider these challenges. For getting the desired accuracy results of cancer type prediction, different efficient approaches should be considered. Various recent studies presented handling the challenges mentioned above.

2.3. Challenges in Cancer prediction research using Machine Learning

2.3.1. High Dimensionality

Due to the massive increase in data dimensions, specifically in somatic genetic mutation datasets for cancer prediction (Zang et al., 2016), the need for more efficient and accurate dimensionality reduction techniques increased. Dimensionality reduction means reducing the features' dimensions by keeping the crucial original features. Various studies addressed the data dimensionality issue to improve cancer prediction process and reduce computational time for training the model. Recent

studies revealed that only a small set of genetic mutations are related to cancer classification.

Different studies proposed cancer type prediction models by implementing feature selection and reduction techniques using the PCA feature selection technique. (Arowol, Abdulsalam, Isiaka, & Gbolagade, 2017), (Alanni, Hou, Azzawi, & Xiang, 2019), and (Adiwijaya, Wisesty, Lisnawati, Aditsania, & Kusumo, 2018). (Arowol et al., 2017) implemented hybrid dimensionality reduction techniques to improve the prediction accuracy. One-Way-ANOVA technique has been implemented as a feature selection to select the most relevant features related to cancer type prediction. PCA and Partial Least Squares (PLS) (Love & McLean, 2011) methods have also been implemented as feature extraction. SVM algorithm has been implemented as a classifier as well as features extraction and reduction techniques to achieve higher accuracy results than implementing an SVM classifier without dimensionality reduction techniques. The proposed model achieved high accuracy results.

Subsequently (Alanni et al., 2019) implemented the same classifier for cancer prediction, SVM classifier using the Microarray cancer dataset. Feature selection methods were also adopted for selecting the most relevant features from the dataset. However, the Leave One Out Cross-Validation technique was implemented to train the model. The developed models considered implementing SVM as a classifier for the dataset.

Similarly, (Adiwijaya et al., 2018) proposed a model for cancer prediction by implementing PCA techniques for feature selection using a DNA Microarray dataset. SVM and Levenberg-Marquardt Backpropagation (LMBP) (Sapna, 2012) algorithms were implemented as classifiers for training the model. SVM classifier is a linear algorithm that separates the classes to find the optimal hyperplane. LMBP is a backpropagation

classifier that aims to find the best weight for the features to improve the classification results. The results demonstrated that LMBP obtained higher accuracy results than the SVM classifier. In this study, different structures for the LMBP classifier were implemented, so implementing a more structured architecture is needed.

Similarly (Sahran, Albashish, Abdullah, Shukor, & Pauzi, 2018) proposed a feature selection model to reduce the high-dimensionality of the tissue components and obtain higher accuracy rates in predicting prostate histopathological images. The model is based on integrating different classifiers and ensemble the results obtained by each classifier. This study has implemented a new approach, which integrates the SVM-Recursive Feature Elimination (SVM-RFE) approach with the Absolute Cosine (AC) filter as a feature selection technique. The approach aims to eliminate the redundancy in the tissue samples. The results revealed that adopting the SVM-RFE-AC (Ramírez-Hernández & Fernandez, 2007) approach in diagnosing cancer can yield higher accuracy rates than the SVM classifier separately.

(Jain & Singh, 2018) adopted an approach similar to (Arowol et al., 2017) by implementing hybrid feature selection methods to reduce the data dimensionality. ReliefF (Urbanowicz, Meeker, La Cava, Olson, & Moore, 2018) and PCA techniques were implemented as feature selection. The study aims to improve the prediction accuracy and reduce the computational time of the model. The proposed model structure starts by implementing the ReliefF feature selection technique as a preprocessing step on the dataset. A list of features is extracted from applying ReliefF to the dataset by assigning a specific weight for each feature. According to a predefined threshold, the list of most essential features is then extracted. PCA method then is applied to the extracted features to be used in the

classification process. The proposed techniques presented high accuracy results when applying a hybrid approach for feature selection. However, the proposed approach can be further improved by testing different recent feature selection techniques.

A study done by (Huljanah, Rustam, Utama, & Siswantining, 2019) developed a Prostate Cancer prediction model using machine learning algorithms. The model is based on an RF classifier along with feature selection techniques. RF determines the number of decision trees to classify cancer with higher accuracy results. The feature selection technique has been applied to reduce the data dimensionality by selecting the prediction process's essential data. The results demonstrate that RF can be used in cancer classification. However, the accuracy levels need to be improved by implementing more advanced feature selection techniques.

(Ghosh, Begum, Sarkar, Chakraborty, & Maulik, 2019) proposed a model for cancer prediction based on implementing feature selection techniques for dimensionality reduction. The proposed technique is Recursive Memetic Algorithm (RMA) (García-Pedrajas, de Haro-García, & Perez-Rodríguez, 2014), which can filter the promising features as biomarkers in predicting cancer type. The RMA implements the feature ranking technique to select the potential biomarkers in the prediction process. The proposed model obtained high accuracy results. However, better results can be revealed by implementing the RMA algorithm using RNA and mRNA sequenced data.

A recent study was done by (Cueto-lópez et al., 2019), assessing different feature selection techniques with different classifiers for predicting colon cancer. RFE-SVM algorithms were adopted in this study. Due to determining the estimated number of features before training the model, a

dropout of some possible relevant features reduced classification accuracy.

Another study done by (Yerukala Sathipati & Ho, 2018) utilized the mRNA sequencing technology in building a model for predicting the stage of breast cancer. The model was developed by implementing the SVM algorithm as a classifier for the cancer stage. Inheritable Bi-objective Combinatorial Genetic algorithm (IBCGA) (Katoch, Chauhan, & Kumar, 2020) was adopted as a feature selection technique. IBCGA is an intelligent algorithm that was widely implemented in solving the optimization problem for vast parameters. The results demonstrated relatively high results in the prediction evaluation.

A recent study done by (Zhao et al., 2019) proposed a model to predict Colorectal Cancer (CRC) based on machine learning algorithms. The proposed model was built by implementing LR and SVM classifiers using the Colon cancer dataset. The model is aimed to classify whether the samples are normal or cancerous. For dimensionality reduction, the grid search algorithm was implemented as a feature extraction technique. Various kernel types, linear, sigmoid, polynomial, and Radial Basis Function, were tested in the prediction process. The results indicate that Radial Basis Function achieved higher results. However, the classification results can be improved by integrating different feature selection techniques like wrapper-based learning to learn the most relevant features during the prediction process.

Feature selection techniques are categorized as wrapper and filter and embedded. Wrapper-based methods depend on the forward selection and backward elimination techniques. In contrast, filter-based methods use statistical calculations and dependence between the dataset's features and classes (Ma & Gao, 2020). Studies showed that wrapper-based feature

selection methods perform better than filter-based methods (Wah, Ibrahim, & Hamid, 2018). Various studies implemented wrapper feature selection techniques in predicting cancer types accurately by eliminating the high dimensionality in the DNA genomic datasets. (Cueto-lópez et al., 2019) conducted a comparative study using machine learning algorithms in predicting colorectal cancer risk. In this study, the SVM-wrapper feature selection technique was applied to rank the most predictive features. Different machine learning classifiers were implemented to evaluate the most efficient and robust classifier in predicting the cancer risk. The classifiers are KNN, SVM, NN, BT, and LR. The results demonstrate that the SVM classifier obtained higher accuracy scores than the other classifiers using feature selection techniques like Recursive Feature Elimination (RFE).

On the other hand, different studies applied a hybrid and integrated approach utilizing the advantages of both wrapper and filter feature selection techniques. (Pavithra, 2017) proposed a model using feature selection technique for predicting cancer type. A Gene expression microarray dataset is used. For reducing the enormous number of features in the dataset, a feature selection approach was implemented. Two steps were performed; the first step is applying filter-based feature selection based on Mutual Information (MI) (Ross, 2014) evaluation to select the most relevant features. The second step is applying wrapper-based feature selection using the Genetic Algorithm (GA) to select the most predictive gene expressions. The results yielded high accuracy results by implementing MI as a filter-based feature selection technique.

(Hancer, 2018) proposed a new approach based on integrated wrapper-filter feature selection techniques. This study aimed to utilize the high performance of wrapper-based feature selection techniques and

computational efficiency of filter-based feature selection techniques. A fuzziness-based method was used to implement a local search approach that deals with linear and discrete data.

Similarly, (Venkataramana, Gracia, Rajavel, & Dodda, 2019) developed a cancer type prediction model based on a hybrid feature selection approach. Parallelized correlation and rank-based feature selection methods were implemented. The hybrid technique aims to reduce and eliminate the irrelevant data from the DNA microarray dataset to increase model accuracy. Parallel classification algorithms are then applied to the selected features to evaluate the performance of predicting cancer type. DT and RF classifiers were applied to evaluate the performance and accuracy of the model. The results revealed that adopting the hybrid approach improves the model's accuracy in selecting the most relevant subset of features and reduces the computational time while preserving the original data. However, different ensemble approaches can be implemented to reduce further the number of features related to cancer prediction.

2.3.2. Class Imbalance

Class Imbalance (CI) is one of the challenges facing researchers in applying machine and deep learning algorithms in different applications due to the datasets' enormous size (Ling & Sheng, 2010). Different studies were conducted handling the CI challenge, specifically in predicting and diagnosing different types of cancers. (Z. Liu, Tang, Cai, Wang, & Chen, 2017) proposed a hybrid approach for predicting cancer type. Weighted Extreme Learning Machine (WELM) algorithm (Zong, Huang, & Chen, 2013) has been implemented. WELM algorithm prevents class imbalance issue by optimizing the feature selection process from the Microarray

dataset. WELM can better improve the classification performance at the algorithmic level by adjusting the input features' parameters. The results demonstrate that implementing the WELM algorithm improves the classification performance. However, different ELM approaches can be implemented and evaluated for better accuracy results. Another study was done by (Tabl, Alkhateeb, ElMaraghy, Rueda, & Ngom, 2019) handling the CI challenge. In this study, a machine learning approach was proposed for predicting breast cancer survival. The model aimed to find the most predictive biomarkers to improve the prediction accuracy. Feature selection maximum Relevance Minimum Distance (mRMD) algorithm (Radovic, Ghalwash, Filipovic, & Obradovic, 2017) was implemented to select the features with a maximum correlation and minimum redundancy in a labeled dataset. The proposed model obtained high accuracy results in predicting the patients' survival from breast cancer.

A recent study was done by (Q. Wang, Zhou, Zhang, Tang, & Chen, 2020), handling the imbalanced data problem in machine learning for pre-prognosis of different cancer types. In the proposed model, Self Paced Learning (SPL) approach was implemented to select the most predictive samples that improve the prediction accuracy results. SPL implements an iterative process on the dataset samples to extract a balanced dataset. The results show that the performance increases when reducing the dataset's imbalanced data as a preprocessing step to improve the classification performance.

On the other hand, (Guo, Liu, Li, & Shang, 2018) proposed a modified model based on Deep Forest (DForest) model called BCDForest. The modified BCDForest model differs from the DForest in that BCDForest trains the multi-class classifiers by implementing a multi-class grained technique to overcome class diversity challenges. Boosting techniques

were applied to select the most relative features in the forests. The results show that the modified model performs well in eliminating the class imbalance challenge.

A more recent study (Shon, Batbaatar, Kim, Cha, & Kim, 2020) proposed a hybrid deep learning model for predicting kidney cancer. The objective of the study is to address the data imbalance issue in small datasets. The cost-Sensitive Hybrid Deep Learning approach was implemented using end-to-end loss function for classifying kidney cancer efficiently. Neural Networks combined with Auto Encoder (AE) (J. Wang et al., 2020) technique was applied with the proposed model to select the non-linear most relevant features to improve the prediction accuracy. The results demonstrate that the proposed hybrid model performs better than the other conventional techniques and improves the prediction accuracy.

(Chang & Chen, 2019) proposed a machine learning-based model for predicting second primary cancers for the patients who survived breast cancer. In this study, the XGBoost algorithm was implemented to predict the risk factors of breast cancer accurately. XGBoost is an optimized gradient boosting algorithm that is based on a decision tree. It can be used to classify different types of applications efficiently. In this study, ensemble learning, transformation, resampling, and clustering have been applied to eliminate the imbalanced data and improve the model performance. However, more advanced feature selection techniques should be adopted to improve the model performance and increase the prediction accuracy.

Similarly, another study (Sherafatian, 2018) proposed a machine learning model for predicting breast cancer. A tree-based algorithm was implemented to select a subtype of features from the miRNA gene expression dataset. The dataset was normalized using the Empirical

Negative Control technique (ENC). The ENC approach is used to control the imbalanced trained data's negative impact. Caret and RF algorithms were applied as a tree-based machine learning technique. Though the proposed model obtained high results. However, more efficient and accurate imbalanced data elimination techniques should be considered.

Feature scaling techniques increase the correlation between the values of the features. A recent study (Sherafatian & Arjmand, 2019) proposed a model for predicting Lung cancer using the LUAD dataset. In their proposed model, a decision tree-based classifier is implemented for predicting cancer status. Remove Unwanted Variation (RUV) normalization method is used in their study for feature scaling. Their proposed model accuracy results may improve more by adopting different normalization techniques.

2.4. Summary

Cancer is a complex disease that results from the abnormal growth of cells. Due to the potential of spreading cancer cells to different parts of the human body, CUPS tumor types may emerge. Determining the site of origin is a challenging task. Hence, multiple studies utilized machine learning technologies along with DNA genomic sequencing technology. Different studies applying various machine learning algorithms using DNA and RNA-Seq genomic sequencing have been discussed in this literature.

In conclusion, different machine learning classifiers obtained high accuracy results of cancer type prediction by utilizing the DNA and RNA genomic sequencing technologies. Subsequently, various pre-processing techniques aim to reduce high dimensionality and class imbalance in the massive size genomic dataset were applied. However,

different hybrid or ensemble models using different feature selection and classification algorithms based on the NGS technique should be considered. According to the literature review study, SVM and RF classification algorithms perform efficiently in DNA and RNA genomic sequencing high dimensional data. The results of the classification can be further improved by utilizing different feature selection techniques.

Hence, our proposed models are pre-processed using efficient feature selection techniques like MI and RFE due to the high accuracy results obtained by utilizing these techniques. MI and RFE feature selection techniques are utilized with the SVM algorithm as a classifier. RF as an embedded feature selection and classifier is also adopted in our proposed models. The proposed models have been developed using NGS RNA-Seq gene expression datasets. The datasets illustrate the expression levels of the genes in terms of up-regulation or down-regulation expression values. These expressions represent the gene mutations that may contribute to the development of different types of cancers.

Chapter 3

The Proposed Approach

3.1. Overview

3.2. Cancer Prediction and Early Detection

3.2.1. Problem Description

3.2.2. The Proposed Models

3.3. Summary

Chapter 3: The Proposed Approach

3.1. Overview

This chapter highlights the proposed models for the prediction and early detection of different types of cancers. Another objective is to find the most potential biomarkers for a specific cancer type that significantly helps in personalized medicine. The proposed models are based on the NGS technique RNA-Seq gene expression using machine learning algorithms. This chapter comprises two parts. The aim of part 1 is to predict and diagnose cancerous or normal individuals from their gene expression values. Part 2 extends the same models to find the most potential biomarkers for three types of cancers.

3.2. Cancer Prediction and Early Detection

3.2.1. Problem Description

Cancer Causes and Types

Cancer is a massive public health problem resulting from mutations of individual genes and abnormal cells development. The DNA inside the cells is comprised of a massive number of individual genes. A set of instructions are transferred to the cells, informing them how to function and divide. In some cases and for different reasons, the cells will ignore the instructions and signals and keep dividing and growing, resulting in cancerous cells (Şirin, Elmas, Seçme, & Dodurga, 2020). The cancerous cells grow and divide uncontrollably and can spread to different parts of the body by destroying the normal tissues. The spread of cancer cells to different body organs is called (Metastatic Cancer). Metastatic cancer

means cancer cells break away from the tumor site they formed initially (Primary Site of Origin). In many cases, diagnosing the primary site of origin with metastatic cancer is problematic, resulting in Cancer of Unknown Primary Site (CUPS) (Rassy, Assi, & Pavlidis, 2020). Diagnosing the primary site of origin helps the physicians in the treatment phase and specializing the medicine per patient. Detecting cancer at early stages reduces the risk and increases the possibility of survival.

DNA

Deoxyribonucleic Acid (DNA) is a molecule consisting of polynucleotides strands that bound around each other, form a double helix. The DNA has four nucleotides, namely Adenine (A), Guanine (G), Thymine (T), and Cytosine (C). These nucleotides are connected as base pairs comprising the organism genome sequence. The nucleotides' base pairs are joined in a chain. The connection occurs according to the pairing rules, where (A with T) and (C with G) forming double-stranded DNA (Jolfaei et al., 2020).

RNA

Ribonucleic Acid (RNA) is a macromolecule that is available in all biological cells. The RNA's role is to transfer the cells' messenger instructions from the DNA and participate in protein synthesis. The DNA carries the genetic instructions required for the growth, cells functions, and life. The chemical structure of DNA and RNA is very similar. However, the RNA has a Uracil (U) base instead of Thymine (T). RNA, in contrast to DNA, is single-stranded, and it has many types. The messenger mRNA is the most crucial type as it carries the instructions from the DNA for protein synthesis. The biological system of the flow of genetic information is called the (Central Dogma). Basically, the system consists of three steps. The first step is DNA replication. The DNA generates multiple copies of

itself by splitting the double strands during the cell division and reconstructing the other single strand to generate a second copy of the DNA. Transcription is the second step in which the mRNA is transcribed or copied from the DNA. The last step occurs when the transcribed mRNA participates in the protein synthesis (Translation) (Sonneveld, Verhagen, & Tanenbaum, 2020), as shown in Figure (3.1).

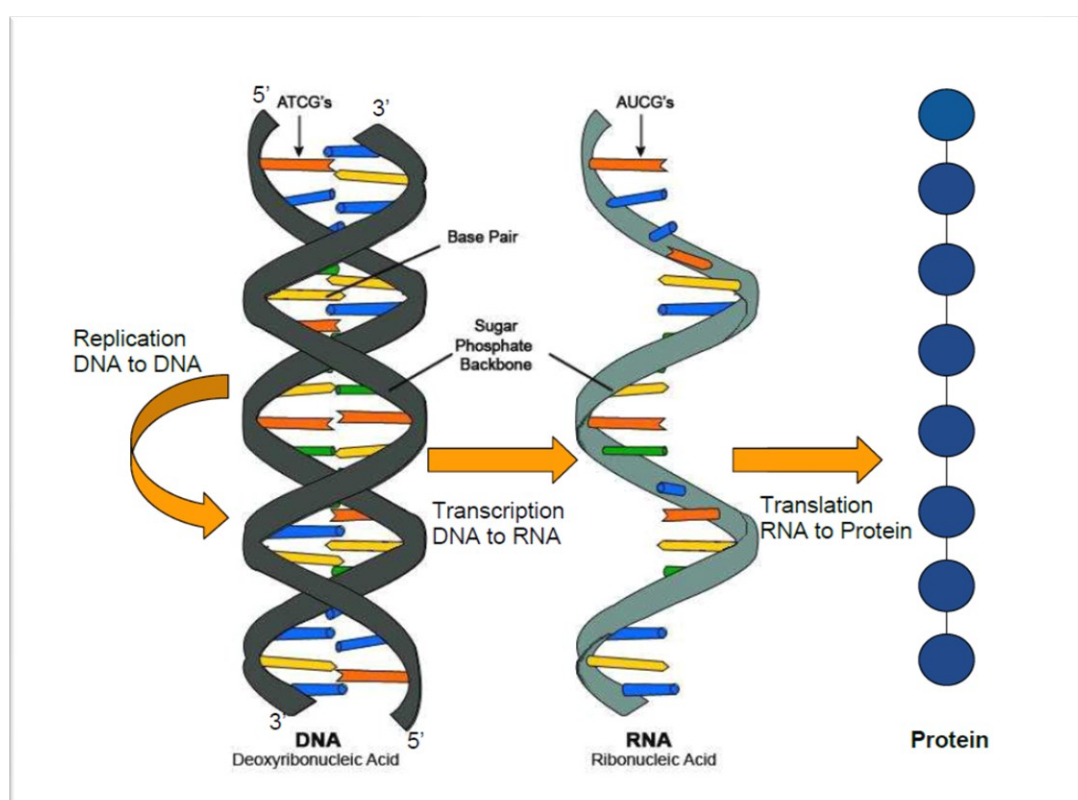


Figure (3.1): Central Dogma system of Molecular Biology (Thakur, 2015)

Mutations

A mutation is a change in the DNA genome sequence of an organism. Mutations can be inherited from parents. This genetic mutation we may have since we were born and can participate in a small percentage of cancers. Other types of mutations can acquire throughout the organism's lifecycle for different reasons. Errors in DNA nucleotides copying during the cell division process, exposure to radiation or chemicals, and

virus infections may lead to mutations in the DNA sequence. Mutations instruct the cells to perform tasks different than the required functions. It may instruct the cells to grow and divide rapidly, resulting in having many cells with the same mutations. Tumor suppressor genes control the tumor cell division and stop their growth. A mutation in the tumor suppressor genes allows the cell to continue growing and dividing in an uncontrolled way leading to cancer formation. Another effect of gene mutation is in the DNA repair genes, which its primary function is to correct any errors that may occur during the DNA copying process. Mutations that occur in the repair genes cause mistakes in DNA repair and leave uncorrected mistakes, leading to cancerous cells (Von Felden et al., 2021).

There are three different types of Mutations, namely insertion, deletion, and substitution. Insertion occurs when more base pairs are added to the DNA sequence resulting in shifting in the sequence, leading to some diseases or health problems. The deletion occurs when one or more base pairs are deleted or missing from the DNA sequence resulting in the protein missing one or more amino acids. Substitution also occurs when some base pairs of the DNA sequence transition to a different sequence causing a change in the organism's DNA sequence. Mutations are the reasons for differences in the organism's characteristics (phenotype) like skin color, hair color, etc. Mutations are not always harmful. Some mutations do not affect the organism's health. On the other hand, some genetic mutations may lead to different health problems and diseases like cancers (Ballhausen et al., 2020).

RNA Sequencing

The transcriptome is the complete array of mRNA transcripts in a specific tissue or cell, including the coding and non-coding. Identifying transcripts

is the main aim of studying transcriptome, where it assists analyzing the functional elements of the genome. Converting the information in the gene into a functional protein is called (Gene Expression). The gene is called up-regulated if it has higher gene expression. On the other hand, it is called down-regulated if the gene has lower gene expression under specific conditions (Liao et al., 2020). Several techniques have been developed to quantify the transcriptome. RNA Sequencing (RNA-Seq) is a developed approach (Imdahl, Vafadarnejad, Homberger, Saliba, & Vogel, 2020). RNA-Seq is mainly based on the Next-Generation Sequencing (NGS) technique, which utilizes Illumina Sequencing platform (Quail et al., 2012).

Different approaches were developed for transcriptome quantification, like microarray hybridization. Microarrays have several limitations in which it does not support in providing information about the RNA expression. It also lacks the quantification quality if the gene was highly or lowly expressed (Shalaby, Fiaschetti, Baumgartner, & Grotzer, 2014).

According to the reasons mentioned above, in this thesis study, the RNA-Seq gene expression technique was adopted as an NGS technique to predict and detect different cancers.

Machine Learning in Bioinformatics

Machine learning involves utilizing the data to learn specific patterns and make predictions based on different computational and mathematical algorithms. Machine learning has been widely used in different applications in bioinformatics due to its computational ability (H. Li et al., 2020). Patterns can be learned from massive datasets and make complex predictions. Different machine learning techniques have recently been

utilized in DNA and RNA genomics sequencing and gene expression(Sharma & Rani, 2021).

3.2.2. The Proposed Models

The proposed models aim to predict and diagnose diverse types of cancers based on the NGS technique RNA-Seq gene expression using machine learning algorithms. Considering the massive size of the RNA-Seq gene expression dataset, several machine learning techniques should be considered. In this study, three machine learning models were developed. An ensemble model was also developed on top of the three models. The methodology of all the proposed models will be explained in more detail in the next section. The general workflow of the proposed models is presented in Figure (3.2). All the developed models follow the same general steps. Starting from the data pre-processing step, which is explained in more detail in the following section. After processing the data, reducing the high dimensionality of the data is an essential step, especially with large datasets. Feature selection techniques are used to achieve this objective. The list of selected features then fed to the machine learning classifiers for training the model. The model performance is evaluated by comparing the predicted results with the actual results of the testing set.

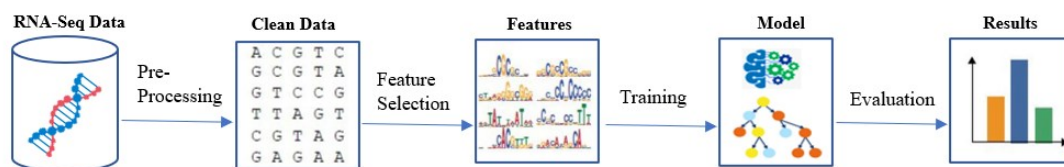


Figure (3.2): General workflow of Machine Learning Model

The detailed steps of the adopted techniques are described in Figure (3.3), and the developed models are explained below in more detail.

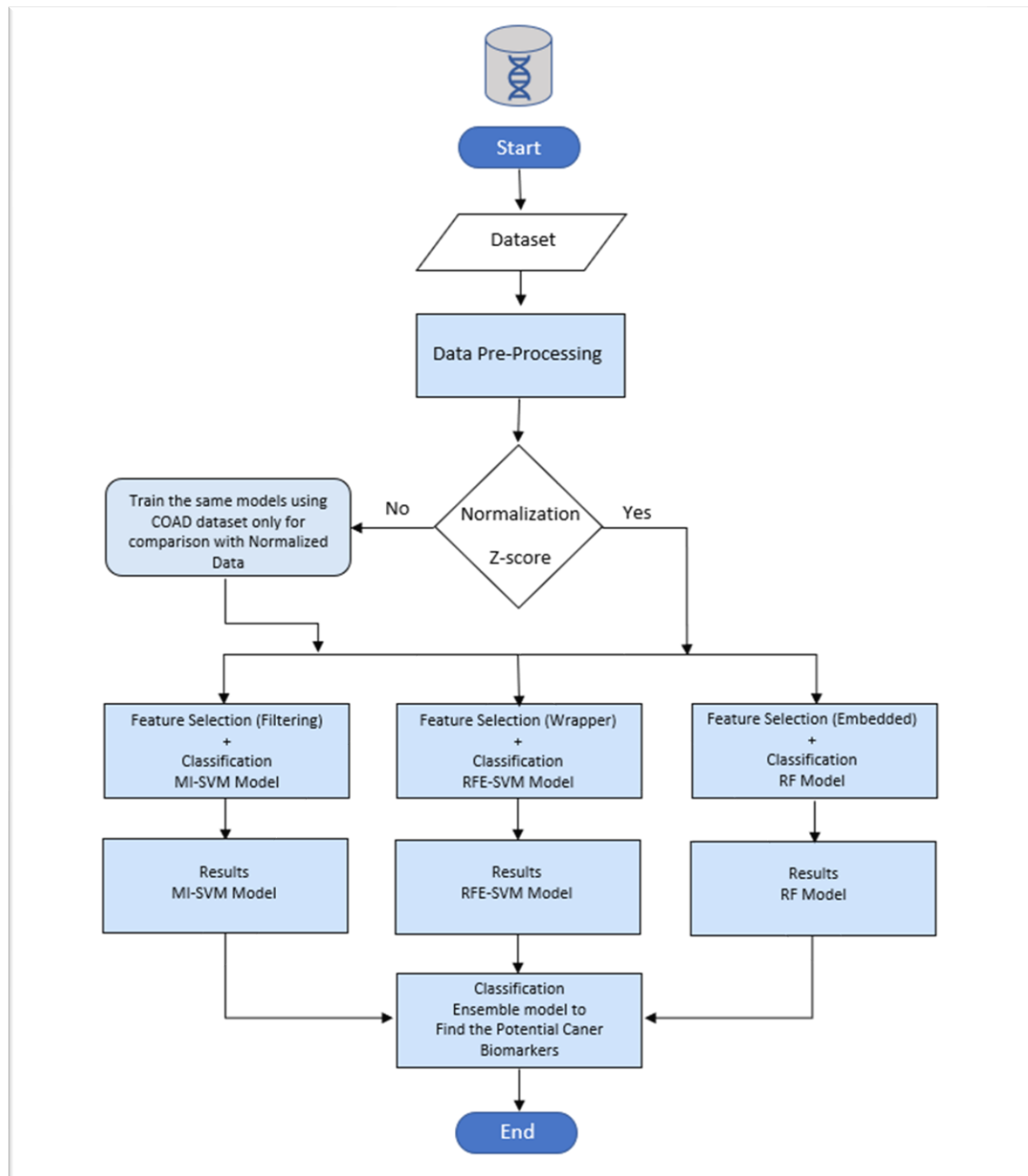


Figure (3.3): The workflow of the proposed Machine Learning Models

Pre-processing

Pre-processing means representing the data in an understandable format. Pre-processing involves several steps. The most critical steps are importing the libraries, importing the dataset, checking the input values, checking the

categorical values, and transforming them to numerical values, splitting the dataset into a training set and testing set. The final step is feature scaling or normalization. Importing the libraries is an essential step in the pre-processing phase. The list of used libraries is summarized in Table (3.1).

Table (3.1): List of imported libraries

Library	Description
import numpy as np	Python package for scientific computing
import pandas as pd	Data Manipulation and Analysis
from sklearn.model_selection import train_test_split	Splitting dataset into training and testing set
from sklearn.preprocessing import StandardScaler	Z-score Data Normalization and standardization
from sklearn.feature_selection import mutual_info_classif	Mutual Information estimation for categorical target values
import time	Required for including time module
from sklearn.svm import LinearSVC	Linear Support Vector Classification
from sklearn.svm import SVC	Support Vector Classification
from sklearn.multiclass import OneVsRestClassifier	Support Vector for multi-class datasets
from sklearn.model_selection import cross_val_score	For importing Cross-Validation library
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score	Printing classification results
import matplotlib.pyplot as plt	Plotting different graphs
from sklearn.feature_selection import RFE	Recursive Feature Elimination for feature ranking
from sklearn.ensemble import RandomForestClassifier	Random Forest Classifier
from PIL import Image	Import the image module for loading images
import warnings	For displaying warning messages
from sklearn.tree import export_graphviz	For plotting and exporting Decision Trees
import pydotplus	Dynamically importing graphs from python
from glob import glob	Retrieving paths from directories or folders
import PIL	For opening and storing images

The datasets were split into training and testing sets for training and performance evaluation purposes.

Cross-Validation is a resampling process used to evaluate machine learning models, especially in datasets with limited samples. There are different types of cross-validation methods. K-fold cross-validation is one of these methods, where the training dataset is split into a k number of folds. In each

fold, part of the training set is used for testing the model and the rest for the training. In this research, 10-fold cross-validation is used.

Feature Scaling

Feature scaling is part of the pre-processing phase. In this step, the values of the features are scaled to a specific range of values. The scaled features' values depend on the implemented feature scaling technique. Z-score, also called a (Standard Score) standardization technique, has been implemented in this study. Z-score is a mathematical measurement that describes the relationship of the value to the mean of a list of values. The range of values of z-score ranges between -3 and +3 standard deviation units. To calculate the z-score value, the mean value μ and the standard deviation σ should be known. Equation (3.1) represents the formula used for calculating z-score values.

$$z = \frac{x - \mu}{\sigma}, \quad \text{Equation (3.1)}$$

where

z represents the distance between the row score and the sample mean measured in units of standard deviation,

x represents the row score,

μ represents the mean of the values,

σ represents the standard deviation of the values.

Normalization in machine learning algorithms is always recommended or even required for algorithms like k-nearest neighbors (X. Wan, 2019). In this study, normalization was implemented in all the proposed models and the used datasets. For evaluation and comparison, the

developed models' performance was measured on one of the selected datasets without implementing normalization (more details are explained in chapter 4). This process aims to evaluate our models' performance and compare the normalized and non-normalized data results.

Feature Selection

Feature selection is an essential step in machine learning techniques, especially with massive, high-dimensional datasets. Feature selection techniques comprise three types: Filter-based, Wrapper-based, and Embedded (Venkataramana et al., 2019). The three techniques were implemented in the developed models. This approach aims to reduce the dataset's high dimensionality and select the most relevant features for the target classes.

Model 1: Mutual Information - Support Vector Machine (MI - SVM)

- **Feature Selection – Filter Based - MI**

MI is a measurement of the mutual dependence of one random variable on another random variable. MI is based on pair random variables measurement. It measures the difference in the joint distribution of the pair variables and the product of the marginal distribution of each one of the pair variables. MI is more general and not limited to a linear dependence like a correlation coefficient. It quantifies the amount of information held by one variable by monitoring the other random variable (Sharmin, Shoyaib, Ali, Khan, & Chae, 2019). MI suffers from a bins problem (divides the data into categories). To avoid this problem, MI-k_neighbors is used. MI is essentially linked to the concept of entropy (a function that quantifies an unpredictable random variable). The formula for calculating the entropy of

MI of two random variables, X and Y , whose joint distribution is defined by $P(X, Y)$, is illustrated in Equation (3.2).

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}, \quad \text{Equation (3.2)}$$

where

X is a random variable that includes a set of values $X = \{x_1, x_2, \dots, x_n\}$,

Y is a random variable that includes a set of values $Y = \{y_1, y_2, \dots, y_n\}$,

$P(X)$ is the probability distribution of X , and the log base is taken to the base of 2.

MI generates non-negative values between two random variables. The value is zero if and only if there is no dependency between the variables. The higher the value, the higher the dependency is. The function of MI depends on the entropy estimation of k-nearest neighbors distances, and it depends on the non-parametric methods (P, H, & Ghaznavi, 2020).

MI has been widely used in bioinformatics and Genomics sequencing studies for selecting the high genetic similarity and relevant coding genes for specific types of diseases (Y. Liu et al., 2020).

The implementation parameters of the MI algorithm are illustrated in Table (3.2). X represents the matrix of features (genes expression), and y represents the target class vector (tumor or normal tissue). `Discrete_features` represent whether the values of the features are continuous or discrete. `N_neighbors` parameter is used to determine the number of neighbors used for MI estimation. On the other hand, the copy

parameter is used to specify whether to copy the initial data during the model training or not. If set to False means, the data will be overwritten.

Random_state parameter is a random value used with the repeated numbers for reproducible results.

Table (3.2): List of MI implementation Parameters

Parameter	Description
x	Features Matrix
y	Target or classes vector
Discrete features	Is data continuous or discrete? Auto = True for sparse x, False for Dense x
n_neighbors	No. of neighbors to use for MI estimation. Default is 3
copy	Whether to copy a given data. True = the data will be copied
random_state	Random value generation for repeated numbers. For reproducible results.

The first proposed model is based on selecting all the features ranked with no zero values that use the Mutual Information (MI) filter-based feature selection technique. The selected features are then fed into the machine learning algorithm to classify cancerous and healthy individuals. Support Vector Machine (SVM) is implemented in this model as a classifier. Detailed information about the developed model is explained in the following section.

- **Classification - SVM**

In the first proposed model, SVM is adopted as a classifier. SVM is a supervised learning algorithm that can be used for classification or regression. It separates different classes by creating hyperplanes in multidimensional space (Platt, 1999). SVM uses an iteration process to minimize the errors by constructing an optimal hyperplane. The largest distance from the hyperplane to the nearest data points of a class is regarded as a good separation. The margin is the space between the support vector points that separate the classes. The larger margin is considered a good margin (Cervantes, Garcia-Lamont, Rodríguez-Mazahua, & Lopez, 2020).

SVM handles the classification of large datasets using different techniques. Radial Basis Function (RBF) kernel has been used in this model. RBF kernel is the default kernel function that can be implemented with non-linear data.

Given n points from a training set of the form $(x_1, y_1), \dots, (x_n, y_n)$, where y_i represents a target class to which the point x_i belongs. y_i can be either 1 or -1.

The hyperplane can be defined as a set of points x given in Equation (3.3).

$$w^T x - b = 0, \quad \text{Equation (3.3)}$$

where

w represents the weight vector to the hyperplane,

$X = (x_1, \dots, x_d)$, set of points in the d -dimensional dataset,

b represents the bias parameter, $\frac{b}{\|w\|}$ represents the offset of the hyperplane to the vector w .

In the case of linear classification, the hyperplane can be described as given in Equation (3.4).

$$y_i(w^T x_i - b) \geq 1, \quad \text{for } 1 \leq i \leq n. \quad \text{Equation (3.4)}$$

The distance between the planes should be maximized. So, the optimization problem is to minimize $\|w\|$ that should satisfy the above Equation.

The w and b values determine the classifier that is given in Equation (3.5).

$$x \rightarrow \text{sgn}(w^T x_i - b), \quad \text{Equation (3.5)}$$

where

$\text{sgn}(\cdot)$ is the signum function.

In the case of non-linear classification, Radial Basis Function (RBF) kernel is used to create non-linear classifiers. RBF function can be computed as given in Equation (3.6).

$$k(\vec{x}_i, \vec{x}_j) = \exp\left(-\gamma \|\vec{x}_i - \vec{x}_j\|^2\right), \quad \text{Equation (3.6)}$$

where

$$\gamma > 0$$

The kernel is subject to the modified $\varphi(\vec{x}_i)$, as given by Equation (3.7).

$$k(\vec{x}_i, \vec{x}_j) = \varphi(\vec{x}_i) \cdot \varphi(\vec{x}_j), \quad \text{Equation (3.7)}$$

w is also in the modified space, as described in Equation (3.8).

$$\vec{w} = \sum_i \alpha_i x_i \varphi(\vec{x}_i). \quad \text{Equation (3.8)}$$

For computing the SVM classifier, the form is given in Equation (3.9).

$$\left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i - b)) \right] + \lambda \|w\|^2, \quad \text{Equation (3.9)}$$

where

λ parameter represents the trade-off between increasing the margin and ensuring that x_i lies on the correct plane.

In the multi-class dataset, the kernel trick is used to learn a non-linear classification, which coincides with linear classification for the modified points $\varphi(x_i)$. Using a kernel function that is given in Equation (3.7) and the vector w that satisfies $w = \sum_{i=1}^n c_i y_i \varphi(x_i)$, where c_i can be obtained using the optimization problem given in Equation (3.10).

$$\begin{aligned}
& f(c_1 \dots c_n) \\
&= \sum_{i=1}^n c_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i c_i (\varphi(x_i) \cdot \varphi(x_j)) y_j c_j, \quad \text{Equation (3.10)} \\
&= \sum_{i=1}^n c_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i c_i k(x_i, x_j) y_j c_j,
\end{aligned}$$

where

$$\sum_{i=1}^n y_i c_i = 0, \text{ and } 0 \leq c_i \leq \frac{1}{2n\lambda} \text{ for all } i.$$

The bias $b = w^T \varphi(x_i) - y_i$,

Accordingly, the classifier can be calculated as $z \rightarrow \text{sgn}(w^T \varphi(z) - b)$.

SVM achieves high accuracy results in many practical applications, especially with high dimensional data (Wu, 2020).

The algorithm flow of the proposed MI-SVM model is explained as follows:

MI-SVM Pseudocode

1	$X: [X]$ All Dataset features
2	Class Label: $Y = [y_1, y_2, y_3, \dots y_i]$
3	Dataset Split:
4	$X_{train}, X_{test}, Y_{train}, Y_{test} \leftarrow \text{train_test_split}(x, y, \text{test_size} = 0.3)$
5	$X_{train}, X_{val}, Y_{train}, Y_{val} \leftarrow \text{train_test_split}(X_{train}, Y_{train}, \text{test_size} = 0.3)$
6	Input:
7	Training Set: $X_{train} = [x_1, x_2, x_3, \dots x_i]$, x_i samples of n-dimensions dataset

```

8  Feature Scaling:
9       $sc \leftarrow \text{StandardScaler}()$ , (scale features using z-score)
10      $X_{train} \leftarrow sc.fit\_transform(X_{train})$ 
11      $X_{test} \leftarrow sc.transform(X_{test})$ 
12  Mutual Information:
13      $mi \leftarrow \text{mutual\_info\_classif}(X_{train}, Y_{train}, \text{discret\_features} = 'auto', n\_neighbors =$ 
14      $3, \text{copy} = 'true')$ , (estimate mutual information between each gene and the target class)
15      $gains \leftarrow \text{array}(mi)$ , (add list of  $mi$  to array)
16      $index \leftarrow (gains != 0)[0]$ , (count no. of genes that ranked by  $mi$  with no zero values)
17  Initialization:
18      $Score\_MI \leftarrow []$ , (accuracy scores list of selected genes)
19  Feature Selection and Classification - Filtered:
20     for  $i$  in range (2 – 1000): (iterates in a loop until reach  $i$  genes)
21          $mi\_features \leftarrow \text{sorted}(\text{range}(\text{len}(gains), \text{key} = \text{lambda } sub: gains[sub]))[-i:]$ ,
22         (sort selected genes in descending order to have the top ranked genes)
23          $x_{train} \leftarrow (X_{train})[:, mi\_features]$  (list of selected genes as a training set)
24          $x_{test} \leftarrow X_{test}[:, mi\_features]$  (list of selected genes as a testing set)
25          $y_{train} \leftarrow Y_{train}[:, ]$  (target class training list)
26          $y_{test} \leftarrow Y_{test}[:, ]$  (target class testing list)
27          $clf \leftarrow \text{SVC}(\text{kernel} = 'rbf')$  (build the model until reach  $i$  genes using SVM rbf
28         kernel for one-cancer datasets), or
29          $clf \leftarrow \text{OneVsRestClassifier}(\text{SVC}(\text{kernel} = 'linear'))$  (build the model until reach
30          $i$  genes using SVM OneVsRest kernel for multi-cancers datasets)
31  Training:
32      $clf \leftarrow clf.fit(x_{train}, y_{train})$ , (training the model)
33  Cross-Validation:
34      $\text{cross\_val\_score}(\text{estimator} = clf, x = x\_train, y = y\_train, cv = 10)$ , (10 Fold CV)
35  Prediction:
36      $y_{pred} \leftarrow clf.predict(x_{test})$ , (predicting the results)
37  Output:
38      $Score\_MI.append(clf.score(x_{test}, y_{test}))$ , (add accuracy scores to a list)

```

The list ($Score_MI$) of most relevant features obtained from the model is stored in a list for further biological analysis to find the most potential biomarkers for different types of cancers (see appendices section).

Model 2: Recursive Feature Elimination - SVM (RFE - SVM)

- **Feature Selection – Wrapper Based – RFE**

Recursive Feature Elimination (RFE) is a wrapper-based feature selection algorithm that implements different machine learning algorithms at the core of the process. RFE fits the model and discards the least essential features until reaching the specified features. The features are ranked using the attributes (`coef_` and `feature_importances_`). The number of features should be specified in RFE (Mohan & Paramasivam, 2020). In some cases, the required number of features is unknown. The model is then fit to reduce further the most relevant genes that obtained the highest accuracy results.

The implementation parameters of the RFE algorithm are illustrated in Table (3.3). `X` represents the matrix of features (gene expression). The estimator parameter represents the classifier used to find the feature importance, SVM in our case. `N_features_to_select` parameter is used to determine the number of features to select for the classification. On the other hand, the `step` parameter is used to specify the number of features to remove from the dataset during the selection process. Finally, the `random_state` parameter is used as a random value used with the repeated numbers for reproducible results.

Table (3.3): List of RFE implementation Parameters

Parameter	Description
<code>X</code>	Features Matrix <code>X</code>
<code>estimator</code>	A linear estimator that can be used to find the feature importance
<code>N_features_to_select</code>	The number of features to select
<code>step</code>	No. of features to remove
<code>random_state</code>	Random value generation for repeated numbers. Used for getting reproducible results.

- **Classification – SVM**

In the second proposed model, the SVM algorithm is used as an estimator. The RFE feature selection process is implemented using the estimator. The estimator trains the model using the training dataset. The purpose of the estimator is to optimize the weights w_i for i number of genes. The iterative process applies backward feature elimination to rank the features. The features with a minor ranking are removed. The process continues until removing all features with the minor ranking. The RFE ranks the features as subsets rather than feature based. Nested feature subset is obtained from this process $F1 \subset F2 \subset \dots \subset F$.

In the case of SVM, the estimation and ranking can be done using the Linear Kernel. Linear Support Vector Classifier (LinearSVC) is used with the linear datasets and makes decisions based on the linear characteristics of the features. In this model, linearSVC has been used with the datasets with one cancer type only (COAD, LUAD, BRCA). In multi-class datasets, the One-versus-Rest (OvR) method is used. This method involves training a single classifier per class. Kernel trick is used in this case to transform the non-linear space to a linear classifier. The equations for calculating the kernel function classification are given in Equation (3.3) to Equation (3.10).

The training process of an OvR classifier is illustrated in the following pseudocode:

Inputs:

- L , training for (Binary classifier algorithm)
- X represents the samples
- labels y where $y_i \in \{1, \dots, P\}$ is the label for the sample X_i

Output:

- f_p for $p \in \{1, \dots, P\}$ (f_p , list of classifiers).

Procedure:

- For each p in $\{1, \dots, P\}$
 - Build a new vector k where $k_i = y_i$ if $y_i = p$ and $k_i = 0$ otherwise
 - Apply L to X, k to obtain f_p

All the classifiers are applied to the testing set as unseen sample x to predict the label p to make decisions for the classifier that obtained the highest accuracy score.

The algorithm flow of the developed RFE-SVM model is explained as follows:

RFE-SVM Pseudocode

1	$X: [X]$ All Dataset features
2	Class Label: $Y = [y_1, y_2, y_3, \dots, y_i]$
3	Dataset Split:
4	$X_{train}, Y_{train}, X_{test}, Y_{test} \leftarrow \text{train_test_split}(X, Y, \text{test_size} = 0.3)$
5	$X_{train}, X_{val}, Y_{train}, Y_{val} \leftarrow \text{train_test_split}(X_{train}, Y_{train}, \text{test_size} = 0.3)$
6	Input:
7	Training Set: $X_{train} = [x_1, x_2, x_3, \dots, x_i]$, x_i samples of n-dimensions dataset
8	Feature Scaling:
9	$sc \leftarrow \text{StandardScaler}()$, (scale features using z-score)
10	$X_{train} \leftarrow sc.\text{fit_transform}(X_{train})$
11	$Y_{test} \leftarrow sc.\text{transform}(Y_{test})$
12	Initialization:
13	$\text{Score_RFE} \leftarrow []$, (accuracy scores list of selected features)
14	Feature Selection and Classification - Wrapper:
15	for i in range $(1 - 1000)$: (iterates in a loop until reach i genes)
16	$\text{clf} \leftarrow \text{LinearSVC}$, (SVM as estimator for one-cancer dataset)
17	$\text{clf} \leftarrow \text{OneVsRestClassifier}(\text{SVC})$, (SVM as estimator for multi-cancers dataset)
18	$\text{selector} \leftarrow \text{RFE}(\text{clf}, n_features_to_select = i, \text{step} = 0.5)$, (build the model until
19	reach i genes)
20	Training:
21	$\text{selector} \leftarrow \text{selector}.\text{fit}(X_{train}, Y_{train})$, (training the model)

```

22 Cross-Validation:
23     cross_val_score(estimator = clf, x = x_train, y = y_train, cv = 10), (10 Fold CV)
24 Prediction:
25      $Y_{pred} \leftarrow selector.predict(X_{test}),$  (predicting the results)
26 Output:
27      $rfe\_features \leftarrow (selector.support\_==1),$  (add RFE features to a list)
28      $Score\_RFE.append(selector.score(X_{test}, Y_{test}),$  (add accuracy scores to a list)
29

```

The list (*Score_RFE*) of most relevant features obtained from the model is stored in a list for further biological analysis to find the most potential biomarkers for different types of cancers (see appendices section).

Model 3: Embedded Random Forest (RF) Model

- **Embedded Feature Selection and Classification – RF**

The third proposed model is based on the embedded feature selection and classification algorithm, Random Forest (RF). RF is an ensemble method that can be used for classification and regression. Multiple decision trees are created using random features during the training process to produce the output class. RF is a bagging algorithm where the dataset is split into different random training sets. However, RF selects a random subset of features for training the individual trees. Hence, there is no correlation or dependency between the trees, which improves the prediction accuracy. RF has been widely used in bioinformatics applications, specifically in DNA and mRNA gene expression and sequencing (Huljanah et al., 2019).

RF uses the general bagging technique as a training algorithm. For a given training dataset $X = x_1, \dots, x_n$ with the target classes $Y = y_1, \dots, y_n$ random samples of the dataset are selected, and repeatedly (S times) train the model and fit trees to these samples. For $s = 1, \dots, S$ training n samples from X , Y ; called x_s, y_s . A classification tree f_s is trained on x_s, y_s . Predictions on

the testing set $\hat{x}$ is made by taking the majority vote given in Equation (3.11).

$$\hat{f} = \frac{1}{S} \sum_{s=1}^S f_s(\hat{x}). \quad \text{Equation (3.11)}$$

RF uses a feature bagging process, which splits a random subset of features during the training cycle. In this process, the features representing solid predictors for the class are selected many times in the S trees.

RF uses the Gini index as an evaluation of the importance of the features. Gini index measures the probability of incorrectly classified features. The values of the Gini index range between 0 and 1, where 0 means all the features belong to one class. On the other hand, 1 implies there is a random distribution of features between the classes. The formula for calculating the Gini index is illustrated in Equation (3.12).

$$\text{Gini Index} = 1 - \sum_{s=1}^S (p_s)^2, \quad \text{Equation (3.12)}$$

where

p_s Represents the probability of a sample classified for a specific class,

n represents the number of classes.

In this implementation, RF is used as a feature selection technique and a classifier based on its embedded characteristic for selecting the most important and relevant features. RF model is then trained using the selected features to predict the classes accordingly.

The list of used parameters in the proposed RF model is illustrated in Table (3.4), where `n_estimators` represent the number of trees used by the model. The `random_state` parameter is used as a random value for the repeated numbers for reproducible results.

Table (3.4): List of RF implementation Parameters

Parameter	Description
<code>n_estimators</code>	No. of trees
<code>random_state</code>	Random value generation for repeated numbers. Used for getting reproducible results.

The algorithm flow of the proposed RF model is explained as follows:

RF Pseudocode

1	$X: [X]$ All Dataset features
2	Class Label: $Y = [y_1, y_2, y_3, \dots y_i]$
3	$features \leftarrow []$, (list of features IDs in the dataset)
4	for i in range ($\text{len}(X)$), (iterate until reach length of X)
5	$features.append(i)$, (save the features IDs in a list)
6	Dataset Split:
7	$X_{train}, Y_{train}, X_{test}, Y_{test} \leftarrow \text{train_test_split}(X, Y, \text{test_size} = 0.3)$,
8	$X_{train}, X_{val}, Y_{train}, Y_{val} \leftarrow \text{train_test_split}(X_{train}, Y_{train}, \text{test_size} = 0.3)$
9	Input:
10	Training Set: $X_{train} = [x_1, x_2, x_3, \dots x_i]$, x_i samples of n-dimensions dataset
11	Feature Scaling:
12	$sc \leftarrow \text{StandardScaler}()$, (scale features using z-score)
13	$X_{train} \leftarrow sc.fit_transform(X_{train})$
14	$Y_{test} \leftarrow sc.transform(Y_{test})$
15	Initialization:
16	$Score_RF \leftarrow []$, (accuracy scores list of selected trees)
17	Feature Selection and Classification - Embedded:
18	for i in range ($1 - 1000$): (iterates in a loop until reach i trees)
19	$clf \leftarrow \text{RandomForestClassifier}(n_estimator = i)$, (build the model until reach i
20	trees)
21	Training:
22	$clf \leftarrow clf.fit(X_{train}, Y_{train})$, (training the model)
23	Cross-Validation:

```

24     cross_val_score(estimator = clf, x = x_train, y = y_train, cv = 10), (10 Fold CV)
25 Prediction:
26      $Y_{pred} \leftarrow clf.predict(X_{test})$ , (predicting the results)
27 Sorting:
28     rf_features ← sorted(zip(map(lambda X: round (x, 4),
29     clf.feature_importances_), features), (sort the the number of genes)
30 Output:
31     Score_RF.append(clf.score(X_test, Y_test), (add accuracy scores to a list)

```

The list (*Score_RF*) of most relevant features obtained from the model is stored in a list for further biological analysis to find the most potential biomarkers for different types of cancers. (see appendices section).

Model 4: Ensemble Model

The fourth implemented model is based on the results of all the previous three models. The main objective of this model is to further refine and filter the most potential biomarkers for different types of cancers. The important features obtained from each previously implemented model are obtained, and the common features between each one of them are stored in a list. SVM classifier is then trained on the obtained list.

The algorithm flow of the developed ensemble model is explained as follows:

Ensemble Model Pseudocode

```

1  X: [X] All Dataset features
2  Class Label: Y = [y1, y2, y3, ... yi]
3  Dataset Split:
4     Xtrain, Ytrain, Xtest, Ytest ← train_test_split (X, Y, test_size = 0.3)
5     Xtrain, Xval, Ytrain, Yval ← train_test_split (Xtrain, Ytrain, test_size = 0.3)
6  Input:
7     mi_features      (list of selected features from MI-SVM model)
8     rfe_features     (list of selected features from RFE-SVM model)
9     rf_features      (list of selected features from RF model)

```

```

10 Initialization:
11      $mi\_rf \leftarrow []$  (list of common genes between MI-SVM and RF models)
12      $rfe \leftarrow rfe\_features$  (list of genes selected by RFE-SVM model)
13 Similar genes between MI-SVM and RF models:
14      $important\_rf \leftarrow []$  (list of RF important eatures ranked with no 0 values)
15     for  $s$  in  $rf\_features$ :
16         if  $s[0] \neq 0$ : (if the gene is used in RF)
17              $important\_rf.append(s[1])$  (add the used genes as a list)
18          $c \leftarrow 0$  (counter of common genes)
19         for  $f$  in  $important\_rf$ : (iterates in a loop until reach  $f$  genes)
20             for  $t$  in  $mi\_features$ :
21                 if  $t == f$ :
22                      $c \leftarrow c + 1$ 
23              $mi\_rf.append(f)$  (add the common genes of MI and RF to a list)
24 Similar genes between MI-SVM, RFE-SVM, and RF models:
25      $similar\_all \leftarrow []$  (list of all similar genes)
26     for  $f$  in  $mi\_rf$ : (iterates in a loop until reach  $f$  genes)
27         for  $t$  in  $rfe$ :
28             if  $t == f$ :
29                  $similar\_all.append(f)$  (add the common genes of MI,RFE, and RF to
30 a list)
31 Dataset Split:
32      $X_{train}, Y_{train}, X_{test}, Y_{test} \leftarrow \text{train\_test\_split}(X, Y, \text{test\_size} = 0.3)$ 
33 Input: Training Set:  $X_{train} = [x_1, x_2, x_3, \dots, x_i]$ ,  $x_i$  samples of n-dimensions dataset
34 Feature Scaling:
35      $sc \leftarrow \text{StandardScaler}()$ , (scale features using z-score)
36      $X_{train} \leftarrow sc.fit\_transform(X_{train})$ 
37      $Y_{test} \leftarrow sc.transform(Y_{test})$ 
38 Classification using similar_all genes:
39      $x_{train} \leftarrow (X_{train})[:, similar\_all]$  (list of all similar genes as a training set)
40      $x_{test} \leftarrow X_{test}[:, similar\_all]$  (list of all similar genes as a testing set)
41      $y_{train} \leftarrow Y_{train}[:, :]$  (target class training list)
42      $y_{test} \leftarrow Y_{test}[:, :]$  (target class testing list)
43      $clf \leftarrow SVC$ , (build the model using SVM)
44 Training:
45      $clf \leftarrow clf.fit(x_{train}, y_{train})$ , (training the model)
46 Cross-Validation:

```

```

47 | cross_val_score(estimator = clf, x = x_train, y = y_train, cv = 10), (10 Fold CV)
48 | Prediction:
49 | y_pred ← clf.predict(x_test), (predicting the results)
50 | Output:
51 | score_all_similar.append(clf.score(x_test, y_test), (add accuracy scores to a list)

```

The list (*score_all_similar*) of the similar genes between all the models is a further reduced list of potential biomarkers for different types of cancers. The list is stored for further biological analysis (see appendices section).

The parameters implemented in this model are the same as the parameters used in the previous models see Table (4.2), Table (4.3), and Table (4.4).

3.3. Summary

This chapter highlights the proposed approach to predict the presence of cancer efficiently. Besides, discovering the most potential biomarkers in the development of different types of cancers. In this study, four models were proposed and implemented, namely MI-SVM, RFE-SVM, RF, and an Ensemble model, combining the results of the three models to reduce the number of genes further and find the most potential biomarkers for cancers. The methodology of implementing the proposed models is explained in detail.

Chapter 4

Experimental Results

4.1. Overview

4.2. Datasets

4.3. The results Analysis and Discussion

4.3.1. Implementation Parameters

4.3.2. Evaluation Criteria

4.3.3. Experimental Results without Feature Scaling

4.3.4. Experiment Results with Feature Scaling

4.3.5. Comparative Study

4.3.6. Biological Analysis of the Potential Biomarkers for Cancers

4.4. Summary

Chapter 4: Experimental Results

4.1. Overview

In this chapter, the research findings and results are highlighted and discussed regarding the aim of the study, which is predicting and classifying cancerous from normal individuals and finding the most potential biomarkers for different types of cancers. The datasets used in this research and their collection details are also explained. The analysis, discussion on proposed models, comparison, and validation are presented. Additionally, this chapter includes several tables and graphs highlighting the findings of the research.

The developed models were encoded in Python Language using cloud-based software (Google Colab) as well as (Anaconda software using Jupyter Notebook 6.0.3). The experiments were implemented using (HP Laptop with Windows 10 operating system, Intel Core™ i7 1.99 GHz, 16 GB RAM), and (HP Workstation with 32 GB RAM, Intel(R) Xeon(R) CPU E5-1620 v4), 3.5 GHz.

4.2. Datasets

In this thesis study, four datasets were used. The datasets are based on RNA-Seq Gene Expression. The gene expression data were sequenced based on the NGS technique. The datasets were collected from The Cancer Genome Atlas (TCGA) program. They can be downloaded from the Genomic Data Commons (GDC) data portal (“GDC Portal,” 2021). The portal contains both open and controlled access datasets. The collected datasets are publicly available with open access. Due to the massive size of the datasets, they are downloaded as separate files. The files were

collected and transformed to a Comma Separated Value (CSV) file using R language. The datasets contain specific samples of individuals with Primary Tumor or Solid Normal Tissue as columns. The gene expressions of each individual are presented as rows containing (60844) expressed genes. Due to the limitation of the maximum number of columns allowed in a CSV file, the data is reversed where the classes are presented as columns, and the genes are presented as rows. For this reason, the data was transposed in the pre-processing step of the models' implementation. The datasets names, description, filtering criteria, and the number of samples available in each are illustrated in Table (4.1). In this study, the proposed models were implemented using three cancer types, namely Colon Adenocarcinoma (COAD), Lung Adenocarcinoma (LUAD), and Breast Carcinoma (BRCA).

Dataset 1 represents the project ID (TCGA-COAD), which includes samples for people with colon cancer as well as normal tissue samples. COAD dataset contains 435 samples in total; 396 are samples with primary colon tumors, and 39 are normal tissue samples. The dataset has been split into training and testing sets during the implementation of the models.

Dataset 2 represents the project ID (TCGA-LUAD), which includes samples for people with lung cancer as well as normal tissue samples. LUAD dataset contains 549 samples in total; 495 are samples with primary lung tumors, and 54 are normal tissue samples.

Dataset 3 represents the project ID (TCGA-BRCA), which includes samples for people with breast cancer as well as normal tissue samples. BRCA dataset contains 1157 samples in total; 1046 are samples with primary breast tumors, and 111 are normal tissue samples.

Dataset 4 combines all the other three datasets where COAD, LUAD, and BRCA datasets are combined in one dataset. As a result, the combined dataset has four classes: colon, lung, breast, and normal tissue samples after mixing all the datasets together. The total number of samples in the combined dataset is 2141, 1937 are samples with primary tumors, and 204 are normal tissue samples.

The datasets have been split into training and testing sets during the implementation of the models. The training set represents 70% of each dataset, which is used to train the model. The testing set represents 30% of each dataset that can be used for testing the model accuracy in predicting the presence of cancer. The testing results are then compared with the actual data to determine the models' accuracy. K-fold cross-validation is also used as a validation criterion for testing the models' skill in predicting new data. In this experiment, $k=10$ is used with k-fold cross-validation. Hence, the number of samples of training and testing sets of each dataset is as follows:

COAD dataset has 304 training samples and 131 testing samples. LUAD dataset contains 384 training samples and 165 testing samples. BRCA dataset has 809 training samples and 348 testing samples. On the other hand, the combined dataset contains 1498 training samples and 643 testing samples.

The filtering criteria for selecting the samples of each dataset from the GDC data portal are illustrated in Table (4.1). The filtering criteria of choosing the samples of each dataset from the GDC portal significantly affect the accuracy of the model in classifying the presence of cancer. Also, the quality of sample sequencing has a significant impact on the accuracy of the results. In this research, the Illumina sequencing platform, one of the

most commonly used platforms for RNA-Seq, has been selected as a sequencing platform for the samples of the datasets. The filtering criteria followed for each dataset are shown as follows:

The selected experimental strategy of all the datasets is RNA-Seq gene expression, representing the sequencing technique that uses the NGS sequencing technology. RNA-Seq reveals the quantity of RNA available in a biological sample at a given time. Similarly, the selected data category of all datasets is transcriptome profiling, which represents the whole set of RNA transcripts (single strand of RNA synthesized from the DNA through the transcription process). The disease type of both COAD and LUAD datasets is adenomas and adenocarcinoma, and Ductal and Lobular neoplasms for the BRCA dataset, which specifies the selected disease type of each tumor.

Table (4.1): Description of the used datasets

	Dataset 1	Dataset 2	Dataset 3	All Datasets Combined
Project ID	TCGA-COAD	TCGA-LUAD	TCGA-BRCA	N/A
Primary Site	Colon	Lung	Breast	Colon+Lung+Breast
Number of Samples	435	549	1157	2141
Primary Tumor Samples	396	495	1046	1937
Normal Tissue Samples	39	54	111	204
Training Samples	70% - 304	70% - 384	70% - 809	70% - 1498
Testing Samples	30% - 131	30% - 165	30% - 348	30% - 643
Filtering Criteria				
Primary Site	Colon	Bronchus and Lung	Breast	Colon + Bronchus and Lung +Breast
Project	TCGA-COAD	TCGA-LUAD	TCGA-BRCA	Combined
Experimental Strategy	RNA-Seq	RNA-Seq	RNA-Seq	RNA-Seq
Data Category	Transcriptome Profiling	Transcriptome Profiling	Transcriptome Profiling	N/A
Disease Type	Adenomas and Adenocarcinomas	Adenomas and Adenocarcinomas	Ductal and Lobular neoplasms	N/A

4.3. The Results Analysis and Discussion

4.3.1. Implementation Parameters

A list of parameters has been used to implement the proposed feature selection and classification techniques in this study. The models were trained by setting the values of the input parameters as illustrated in Table (4.2), Table (4.3), and Table (4.4), respectively. The description of all the parameters used in the models' implementation is explained in detail in chapter 3 (Section 3.2.2.). The values of the MI-SVM model parameters are as follows: x and y values are set to the X_train and y_train to represent the training and testing sets, respectively. Discrete_features is set to auto, n_neighbors is set to 3 neighbors, copy parameter is set to True, and random_state value is 42 as a random value.

Table (4.2): List of MI-SVM model implementation Parameters Values

Parameter	Value	Description
x	X_train	Features Matrix
y	y_train	Target or classes vector
Discrete_features	auto	Is data continuous or discrete? Auto = True for sparse x, False for Dense x
n_neighbors	3	No. of neighbors to use for MI estimation. Default is 3
copy	True	Whether to copy a given data. True = the data will be copied
random_state	42	Random value generation for repeated numbers. For reproducible results.

The RFE-SVM model parameters' values are set as follows: X_train represents the features matrix. The estimator parameter value is set to either LinearSVC, RBF to train the COAD, LUAD, and BRCA datasets or to OneVsRestClassifier to train the multi-class combined dataset. Finally, the random_state value is set to 42 as a random value for reproducible results.

Table (4.3): List of RFE-SVM model implementation Parameters Values

Parameter	Value	Description
x	X_train	Features Matrix
estimator	LinearSVC, RBF – with OneVsRestClassifier for multi-class dataset	LinearSVC: A linear estimator that can be used to find the feature importance. RBF: for non-linear data.
N features to select	<i>i</i> (based on the loop)	The number of features to select
step	0.5	No. of features to remove
random_state	42	Random value generation for repeated numbers. Used for getting reproducible results.

The *n_estimators* value of the RF model is set to *i*, where the number of trees selected by the model can be determined according to the highest accuracy obtained during the model training.

Table (4.4): List of RF model implementation Parameters Values

Parameter	Values	Description
n_estimators	<i>i</i> (based on the loop)	No. of trees
random_state	42	Random value generation for repeated numbers. Used for getting reproducible results.

4.3.2. Evaluation Criteria

As an evaluation of the models' performance, confusion matrix results are used to analyze how the model is confused in predicting each class. Accuracy, precision, recall, and F1 metrics were used as evaluation metrics for the models' performance. The formulas for calculating the classification metrics are illustrated in Equation (4.1), Equation (4.2), Equation (4.3), and Equation (4.4), respectively.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad \text{Equation (4.1)}$$

where

TP = True Positive (predicted as cancerous and it is actually cancerous)

TN = True Negative (predicted as normal and it is actually normal)

FP = False Positive (predicted as cancerous and it is actually normal)

FN = False Negative (predicted as normal and it is actually cancerous)

$$Precision = \frac{TP}{TP+FP} \quad \text{Equation (4.2)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{Equation (4.3)}$$

$$F1 = 2 * \frac{(Recall*Precision)}{Recall+Precision} \quad \text{Equation (4.4)}$$

The precision score of a class represents the number of positive predictions truly related to the positive class. The recall metric measures how the positive predictions are truly related to the positive samples. The last metric, the F1 measure, balances the precision and recall metrics' scores in the dataset.

4.3.3. Experimental Results without Feature Scaling

The first three developed models' (MI-SVM, RFE-SVM, RF) primary objective is to predict cancer probability and classify primary tumors from normal solid tissues efficiently with high accuracy results. The fourth proposed model's (Ensemble model) primary objective is to further reduce the selected biomarkers (genes) in the first three models and select the most correlated genes for developing a specific cancer type. Four datasets were applied to each model, namely COAD, LUAD, BRCA, and the combined dataset, as explained in (section 4.2).

In the first experiment, the proposed models were applied to the COAD dataset only, in which no feature scaling was done on the dataset in the pre-processing step. The models' implementation using non-scaled features is

not focused on in this thesis study. This experiment aims to compare the models' classification results without perform feature scaling. Although the classification accuracy results were high, as illustrated in Table (4.5), which was around 99%, the models could not find the most common genes between them for the COAD dataset. The reason is that normalized gene expression values increase the correlation of RNA-Seq within tissues of the up-regulated and down-regulated expressions in genes. Because no common genes between all the three models in this implementation were found, no classification was done. As a result, this does not achieve the second objective of the study, finding the most correlated genes contributing to cancer development. For this reason, the implementation of the models without perform feature scaling was not applied on the other datasets.

Table (4.5): Accuracy results of COAD Dataset without Feature Scaling

Dataset	MI-SVM %	RFE-SVM %	RF %	Ensemble Model %
COAD	98.8	99.2	99.4	No common Genes found, so no classification done

4.3.4. Experiment Results with Feature Scaling

- **Model 1: MI-SVM Model**

The accuracy scores of the MI-SVM model are illustrated in Figure (4.1), Figure (4.2), Figure (4.3), and Figure (4.4). The graphs illustrate the accuracy scores of the 1000 iterations of selecting the top relative genes to each type of cancer according to the used dataset. The x-axis represents the number of genes, and the y-axis represents the accuracy scores obtained for the top selected genes. For instance, in Figure (4.1), for the COAD dataset, at the number of genes around 2, the accuracy score was 100%. The same with the number of genes from around (10-400), the accuracy

score was 100%. This result can be interpreted as these genes are highly contributing to colon cancer. The same interpretation is for all MI-SVM model four graphs. The graph scores fluctuate or decrease at some points, which interprets that the gene contributes negatively to that cancer type as the gene expression levels are low.

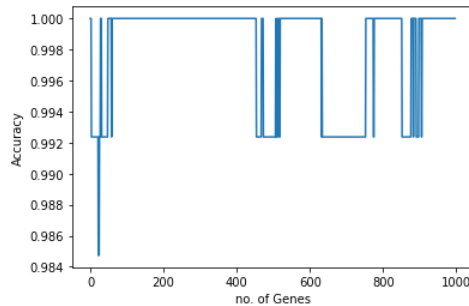


Figure (4.1): Accuracy of MI-SVM model in COAD Dataset

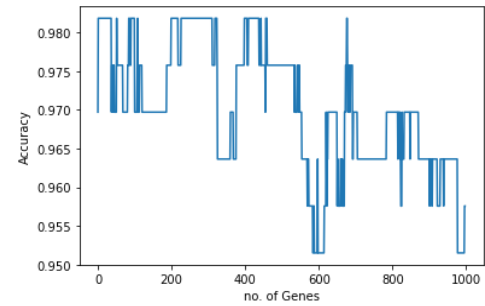


Figure (4.2): Accuracy of MI-SVM model in LUAD Dataset

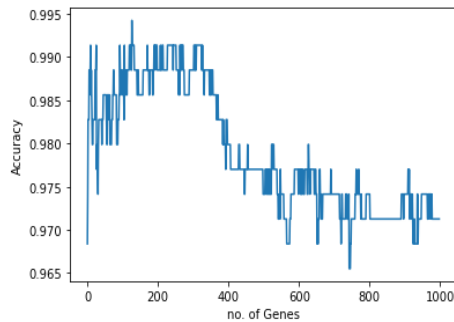


Figure (4.3): Accuracy of MI-SVM in BRCA Dataset

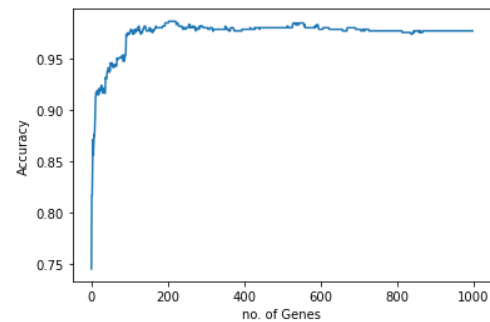


Figure (4.4): Accuracy of MI-SVM model in Combined Dataset

Table (4.6) illustrates the MI-SVM model confusion matrix and classification report of the four used datasets. The table includes the confusion matrix values and the validation metrics, namely accuracy, precision, recall, and F1 scores. As illustrated, MI-SVM achieved high accuracy results for all the datasets in an average of (99-100%).

The model could efficiently select the most contributing genes to colon cancer development with 100% accuracy in the COAD dataset. The results

were validated using the unseen testing set, representing 30% of the dataset. The COAD dataset achieved the highest accuracy results due to the reason that the gene expression of this dataset is more reliable and expressed accurately, and the data correlation is high. It can also be explained that the quality control that takes place over the sequenced fragments during the RNA sequencing and synthesizing process is more accurate than the other datasets.

The model also achieved high accuracy results in the BRCA and LUAD datasets of 99.42% and 98.18%, respectively. As illustrated in the confusion matrix, the model could predict 314 breast tumor samples out of 315 and 32 normal samples out of 33 with an accuracy of 99.42% in the BRCA dataset. The precision and recall metrics results are considerably high. In the LUAD dataset, the model could predict all the tumor samples (149) correctly, which can be noticed from the recall value (100%), proving that the model could correctly identify the actual tumor samples. However, the normal tissue's recall score is 81% because the model could not correctly predict three normal samples out of 16. The reason for that could be because of the small number of normal samples in the dataset.

For evaluating the model's performance, the classification results have been examined on a multi-class dataset by combining all the three datasets, COAD, LUAD, BRCA, under one dataset (combined dataset). The combined dataset consists of four classes, namely colon, breast, lung, and normal tissue. As illustrated in Table (4.6), the model accurately classified different types of cancers with 99% accuracy, proving that the model is efficient in classifying cancerous samples even with multi-cancers data, which is challenging to be identified since some genes are highly contributing to two or more cancer types. Some are contributing to one type but not to the other. The precision scores of all tumors in the combined

dataset are high ranging from 99-100%, which means the model could identify tumor samples correctly as the model identified 117 colon tumors out of 119, 147 lung tumors out of 149, and 311 breast tumors out of 314. However, the precision score of the normal tissue samples is 94%. The model could identify 59 normal tissue out of 61. The percentage is less compared with the tumor precision due to the number of normal samples in the combined dataset compared to the tumor samples.

Table (4.6): MI-SVM Model Classification Results Report

MI-SVM Model Confusion Matrix						
Dataset		TP	TN	FP	FN	
COAD		119	12	0	0	
LUAD		149	13	3	0	
BRCA		314	32	1	1	
COMBINED	Actual		Predicted			
			COAD	LUAD	BRCA	NORMAL
		COAD	117	2	0	0
		LUAD	0	147	1	1
		BRCA	0	0	311	3
		NORMAL	0	1	1	59
MI-SVM Model Classification Report						
	Accuracy	Precision	Recall	F1-score	Support	
COAD	1.00	1.00	1.00	1.00	119	Tumor
		1.00	1.00	1.00	12	Normal
LUAD	0.98	0.98	1.00	0.99	149	Tumor
		1.00	0.81	0.90	16	Normal
BRCA	0.99	1.00	1.00	1.00	315	Tumor
		0.97	0.97	0.97	33	Normal
COMBINED	0.99	1.00	0.98	0.99	119	Colon
		0.99	0.99	0.99	314	Breast
		0.99	0.99	0.99	149	Lung
		0.94	0.97	0.95	61	Normal

- **Model 2: RFE-SVM Model**

Figure (4.5), Figure (4.6), Figure (4.7), and Figure (4.8) illustrate the accuracy scores of the RFE-SVM model. The graphs illustrate the accuracy scores of the 1000 iterations of selecting the top relative genes to each type of cancer according to the used dataset. The x-axis represents the number of genes, and the y-axis represents the accuracy scores obtained for the top

selected genes. The exact interpretation of MI-SVM is reflected in the below graphs, where the top correlated genes obtain higher accuracy scores. The decreasing scores in the BRCA dataset can be interpreted as the model detected the most correlated features early. The other accuracy scores determine the contribution level of the genes to cancer. The same interpretation is for all RFE-SVM model four graphs. The graph scores fluctuate or decrease at some points, which means that the gene contributes negatively to that cancer type as the gene expression levels are low.

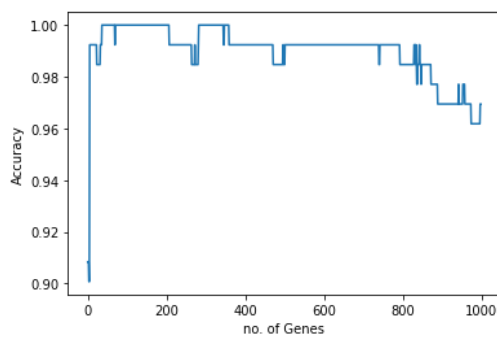


Figure (4.5): Accuracy of RFE-SVM in COAD Dataset

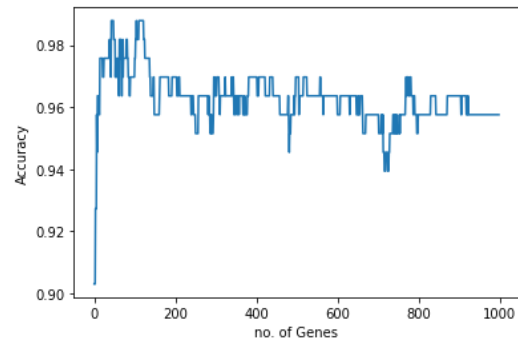


Figure (4.6): Accuracy of RFE-SVM in LUAD Dataset

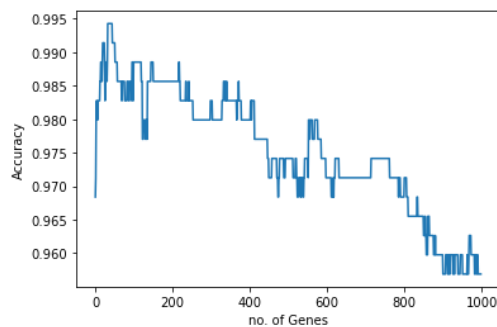


Figure (4.7): Accuracy of RFE-SVM model in BRCA Dataset

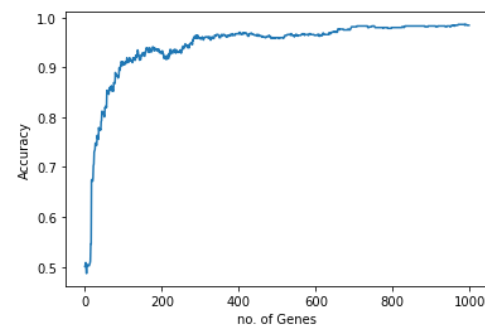


Figure (4.8): Accuracy of RFE-SVM model in Combined Dataset

Table (4.7) illustrates the RFE-SVM model confusion matrix and classification report of the four used datasets. The table includes the accuracy, precision, recall, and F1 scores. As illustrated, RFE-SVM

achieved high accuracy results for all the datasets in an average of (98-100%). The precision and recall metrics results are considerably high. However, the combined cancer dataset scores less in predicting the normal tissue due to the small number of normal samples in the dataset.

The model could efficiently select the most contributing genes to colon cancer development with 100% accuracy in the COAD dataset. The results were validated using the unseen testing set, representing 30% of the dataset. The COAD dataset achieved the highest accuracy results due to the reason that the gene expression of this dataset is more reliable and expressed accurately, and the data are highly correlated. It can also be explained that the quality control that takes place over the sequenced fragments during the RNA sequencing and synthesizing process is more accurate than the other datasets.

The model also achieved high accuracy results in the LUAD and BRCA datasets of 98.78% and 99.42%, respectively. As illustrated in the confusion matrix, the model could predict 148 lung tumor samples out of 149 and 15 normal samples out of 16 with an accuracy of 98.78% in the LUAD dataset, which can be noticed from the recall value (99%), proving that the model could correctly identify the actual tumor samples. However, the normal tissue's recall score is 94% because the model could not correctly predict one normal sample out of 16. The reason for that could be because of the small number of normal samples in the dataset. In the BRCA dataset, the model could predict 314 breast tumor samples out of 315 and 32 normal tissue samples out of 33 correctly with an accuracy of 99%.

For evaluating the model's performance, the classification results have been examined on a multi-class dataset by combining all the three datasets, COAD, LUAD, BRCA, under one dataset (combined dataset). The

combined dataset consists of four classes, namely colon, breast, lung, and normal tissue. As illustrated in Table (4.7), the model accurately classified different types of cancers with 98% accuracy, proving that the model is efficient in classifying cancerous samples even with multi-cancers data, which is challenging to be identified since some genes are highly contributing to two or more cancer types. Some are contributing to one type but not to the other. The precision scores of all tumors in the combined dataset are ranging from 97-100%, which means the model could identify tumor samples correctly as the model identified 118 colon tumors out of 119, 148 lung tumors out of 149, and 304 breast tumors out of 314. The model identified eight tumor samples as normal tissues. Hence, the normal tissue samples' precision score is 88%, representing how the model could identify the normal tissues that were truly normal. The percentage is less compared with the tumor precision due to the number of normal samples in the combined dataset compared to the tumor samples. While the normal tissue samples' recall score in the combined dataset is 100% represents the actual normal tissues correctly identified.

Table (4.7): RFE-SVM Model Classification Results Report

RFE-SVM Model Confusion Matrix							
Dataset			TP	TN	FP	FN	
COAD			119	12	0	0	
LUAD			148	15	1	1	
BRCA			314	32	1	1	
COMBINED	Actual		Predicted				
			COAD	LUAD	BRCA	NORMAL	
		COAD	118	1	0	0	
		LUAD	0	148	0	1	
		BRCA	1	2	304	7	
		NORMAL	0	0	0	61	
RFE-SVM Model Classification Report							
	Accuracy	Precision	Recall	F1-score	Support		
COAD	1.00	1.00	1.00	1.00	119	Tumor	
		1.00	1.00	1.00	12	Normal	
LUAD	0.99	0.99	0.99	0.99	149	Tumor	
		0.94	0.94	0.94	16	Normal	
BRCA	0.99	1.00	1.00	1.00	315	Tumor	

		0.97	0.97	0.97	33	Normal
COMBINED	0.98	0.99	0.99	0.99	119	Colon
		1.00	0.97	0.98	314	Breast
		0.98	0.99	0.99	149	Lung
		0.88	1.00	0.94	61	Normal

- **Model 3: RF Model**

Figure (4.9), Figure (4.10), Figure (4.11), and Figure (4.12) illustrate the accuracy scores of the RF model. The graphs illustrate the accuracy scores of the 1000 iterations of the selected number of trees, which contains the most important features in classifying a specific cancer type. The x-axis represents the number of trees, and the y-axis represents the accuracy scores obtained for the top selected trees. The graphs represent the accuracy scores obtained for each selected number of trees. As illustrated in the graphs, RF achieved high accuracy results for all the used datasets.

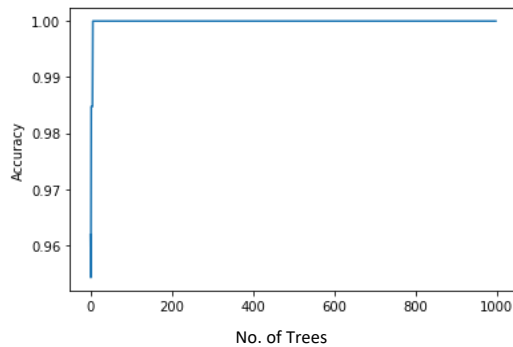


Figure (4.9): Accuracy of RF model in COAD Dataset

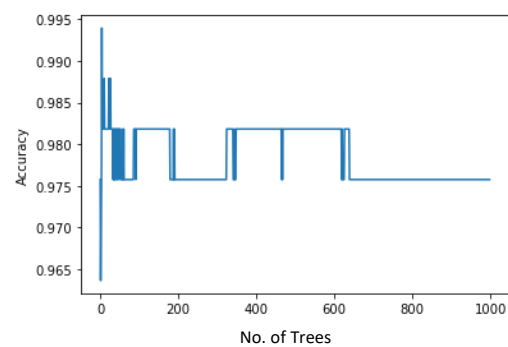


Figure (4.10): Accuracy of RF model in LUAD Dataset

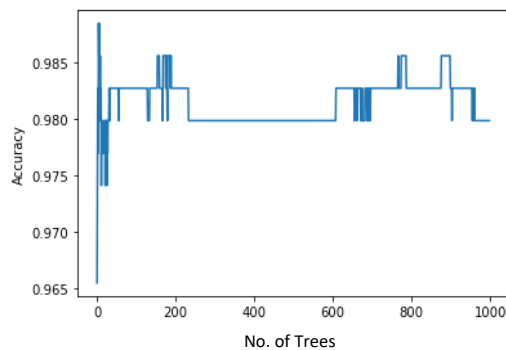


Figure (4.11): Accuracy of RF model in BRCA Dataset

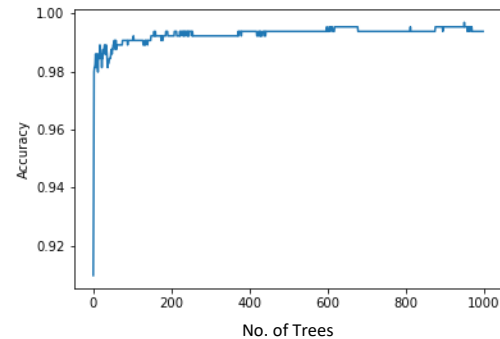


Figure (4.12): Accuracy of RF model in Combined Dataset

Table (4.8) illustrates the RF model confusion matrix and classification report of the four used datasets. The table includes the accuracy, precision, recall, F1 scores. As illustrated, RF achieved high accuracy results for all the datasets in an average of (99-100%). The precision and recall metrics results are considerably high. RF model could efficiently classify the tumor samples as illustrated in the confusion matrix. Only 1 out of 16 normal tissue samples is predicted as a lung tumor in the LUAD dataset. As well, 2 out of 33 normal samples are predicted as breast tumors, and 2 out of 315 breast tumor samples are predicted as normal due to the small number of normal samples in the dataset. The RF model performed efficiently in classifying tumor and normal tissue samples even in the combined dataset with four classes with an accuracy score of 99.68%. As illustrated, the precision and recall scores are 100% of the tumor cancers (colon, lung, breast), while it reached 98% for the normal tissue due to the number of samples compared to the other classes. The results prove that the RF model achieved the highest and more stable results between the three implemented models.

Table (4.8): RF Model Classification Results Report

RF Model Confusion Matrix							
Dataset			TP	TN	FP	FN	
COAD			119	12	0	0	
LUAD			149	15	1	0	
BRCA			313	31	2	2	
COMBINED	Actual		Predicted				
			COAD	LUAD	BRCA	NORMAL	
		COAD	119	0	0	0	
		LUAD	0	149	0	0	
		BRCA	0	0	313	1	
		NORMAL	0	1	0	60	
RF Model Classification Report							
	Accuracy	Precision	Recall	F1-score	Support		
COAD	1.00	1.00	1.00	1.00	119	Tumor	
		1.00	1.00	1.00	12	Normal	
LUAD	0.99	0.99	1.00	1.00	149	Tumor	
		1.00	0.94	0.97	16	Normal	
BRCA	0.99	0.99	0.99	0.99	315	Tumor	

		0.94	0.94	0.94	33	Normal
COMBINED	0.99	1.00	1.00	1.00	119	Colon
		1.00	1.00	1.00	314	Breast
		0.99	1.00	1.00	149	Lung
		0.98	0.98	0.98	61	Normal

Table (4.9) illustrates and summarizes the obtained results from all the proposed models using four datasets. As shown in the table, the proposed models' accuracy results in predicting the presence of cancer achieved high results using the most important genes that contribute to the development of different types of cancers. The models' performance was evaluated on three datasets with one cancer type (COAD, LUAD, BRCA). The fourth dataset combined all three datasets, and it has four classes (colon, lung, breast, normal tissue). The results obtained from the combined dataset were also high, which implies that the model performs well with multi-class datasets.

Table (4.9): Accuracy results using top selected genes

Dataset	MI-SVM %	RFE-SVM %	RF %	Ensemble Model %
COAD	100	100	100	99.23
BRCA	99.42	99.42	98.85	98.21
LUAD	98.18	98.78	99.39	93.33
COMBINED	98.60	98.60	99.68	97.00

• **Model 4: Ensemble Model**

The fourth implemented model is an ensemble model of the three models (MI-SVM, RFE-SVM, RF). Implementing this model aims to find the top potential biomarkers for the development of different types of cancers. In this model, the list of the common genes between MI-SVM, RFE-SVM, and RF models is used to classify the tumor and normal tissues. The ensemble model performed well with high accuracy results. The model was trained on significantly reduced features, see Table (4.10). For instance, the COAD dataset has five top common genes reduced from 60844 genes, and it is achieved an accuracy of 99.23%, which proves the selected features

have a significant correlation to colon cancer. A biological analysis was done on these genes for further analysis to ensure that the top selected features are significant biomarkers for colon cancer. The results of this analysis are illustrated in (section 4.3.6.). The lists of the top selected potential biomarkers obtained from all the datasets are added to the (Appendices section).

Table (4.10): Number of top selected genes

Dataset	MI-SVM	RFE-SVM	RF	Ensemble Model
COAD	30	50	165	5
BRCA	129	35	64	2
LUAD	38	43	24	1
COMBINED	206	999	2504	52

The models' performance has been examined using all the genes in the datasets without implementing feature selection techniques. Through the results shown in Table (4.11) and compared to the obtained results after implementing feature selection techniques Table (4.9), it is evident that the classification results have improved significantly. The results confirm the necessity of adopting feature selection techniques in the classification, especially in the datasets of high dimensionality.

Table (4.11): Accuracy results using all the genes–No Feature Selection

Dataset	SVM %
COAD	97
BRCA	92
LUAD	94
COMBINED	95

Table (4.12) illustrates the implementation processing time of all the proposed models. The time varies for each dataset according to the number of samples in each one. As shown, the first model (MI-SVM) has less processing time than the other models. The second model (RFE-SVM) has the longest processing time due to the algorithm's iterative nature. Besides, with the combined dataset, which has multi-class, the OneVsRest

technique is used within the RFE algorithm. This technique needs an extended processing time while splitting the classes with each run. The third model (RF) processing time is ranging between the other two models.

Table (4.12): Implementation Processing Time of all the Models in Seconds

Dataset	MI-SVM	RFE-SVM	RF
COAD	93	6092	4000
BRCA	138	11202	8510
LUAD	113	6722	4981
COMBINED	689	120654	36205

4.3.5. Comparative Study

A comparative study has been conducted with other recent studies in cancer prediction using DNA and RNA sequencing technology based on different machine learning algorithms. According to the results shown in Table (4.13), the proposed models achieved high and efficient results, proving that several factors could affect the classification accuracy. Starting from collecting the datasets and the filtering criteria when downloading the data. The filtering criteria used in the selected datasets are illustrated in Table (4.1).

Another factor that may affect the accuracy results is data cleaning and pre-processing. Feature scaling or normalization is part of the data pre-processing phase, where the features' values are transformed to a form with a specific range. The models are trained on this range values, increasing the correlation of RNA-Seq within tissues of the up-regulation and down-regulation expressions in genes. In the proposed models, the z-score normalization technique is used, as explained in Chapter 3. A recent study (Sherafatian & Arjmand, 2019) proposed a model for predicting Lung cancer using the LUAD dataset. In their proposed model, a decision tree-

based classifier is implemented for predicting cancer status. Remove Unwanted Variation (RUV) normalization method is used in their study for feature scaling. The accuracy score of their proposed model reached 91.2%, whereas using z-normalization implemented in our proposed models, the accuracy results were higher, around 99% using the same LUAD dataset.

On the other hand, feature selection is also an essential step to reduce the number of features to the maximum number of features that may contribute to the target class's prediction. Feature selection is an essential step, especially with high-dimensional data. In the RNA-Seq gene expression datasets, the number of expressed genes is exceptionally high (60844). The selection of feature selection techniques with the proper classification algorithm also affects the accuracy rate in training the model and predicting the target class accurately. For instance, (Zhao et al., 2019) implemented LR as a feature selection method with SVM as a classifier for predicting colorectal cancer using the colon dataset. The accuracy result of the proposed method was 91% for predicting colon cancer. LR is limited to the linear relationships between features and needs a large sample size of the dataset, affecting the accuracy scores of the model. MI and RFE feature selection techniques work efficiently in selecting the most correlated genes that may contribute as a biomarker for cancer. Their competence in selecting the most correlated features to the target class has been reported in different studies (Zou et al., 2019), (Lin, Li, Zhang, Su, & Wei, 2018). A recent study was done by (Cueto-lópez et al., 2019), assessing different feature selection techniques with different classifiers for predicting colon cancer using the colon cancer dataset. The accuracy result of the proposed model reached around 69% using RFE-SVM algorithms due to determining the estimated number of features before training the model

resulting in a drop out of some possible relevant features and reducing the classification accuracy. In our proposed models, an iterative process is implemented for selecting the most relevant features.

Several studies adopted deep learning techniques in the classification of different cancer types. (Xiao et al., 2018b) implemented Stacked Sparse Auto Encoder (SSAE) technique for predicting breast cancer using the BRCA dataset, where Auto Encoder is used for feature selection. The result of the proposed model was 96.23%. Comparing the results with our proposed models using the same BRCA dataset confirms the models' efficiency in tumor classification and finding the most potential biomarkers for cancer. On the other hand, deep learning techniques are characterized by their high ability to select the related features. Still, it requires a high processing time and a large number of samples.

According to the state-of-the-art studies, not all studies considered all the factors mentioned above when building the model, affecting classification accuracy. The results illustrate that the proposed models outperform different machine and deep learning techniques. The proposed models optimized the processing time compared to deep learning techniques, which require a long processing time, especially with massive size datasets like DNA sequencing and RNA-Seq gene expression.

Table (4.13): Comparison of the Proposed Models with other Studies

Study	Publication year	Method	Accuracy Scores (%)
(Sherafatian & Arjmand, 2019)	2019	RUV Normalization - Decision Tree	±91.2
(Zhao et al., 2019)	2018	Logistic Regression-SVM	±91
(Cueto-lópez et al., 2019)	2019	Wrapper-RFE-SVM	±68.4
(Xiao et al., 2018b)	2018	Sparse Stacked Auto Encoder	±96.23
This Thesis Study	N/A	Proposed MI-SVM model	±99.2
This Thesis Study	N/A	Proposed RFE-SVM	±99.4
This Thesis Study	N/A	Proposed RF model	±99.4

4.3.6. Biological Analysis of the Potential Biomarkers for Cancers

As explained in the experimental results, using the ensemble model that combines all the top selected genes from all the developed models, a list of common potential biomarkers for specific types of cancers is collected. As illustrated in Table (4.9), the top common genes selected for colon cancer were five. A biological study has been done on these genes to confirm whether the proposed models can efficiently discover the most correlated biomarkers for a specific type of cancer. The genes' names are (BEST4, VSTM2A, LINC00974, SPIB, and CA7). BEST4 (“Tissue Expression of BEST4 - Staining in Colon - The Human Protein Atlas,” 2021) is a member of the Bestrophin gene family, transmembrane proteins. Many studies were done by (Xu et al., 2017) and (Zhu et al., 2019), approving that the BEST4 gene is predominantly expressed in colon cancer.

VSTM2A gene is a V-set and transmembrane domain-containing protein 2A (“Expression of VSTM2A in Colorectal Cancer - The Human Protein Atlas,” 2021). It was approved that the VSTM2A gene is a novel antagonist potential suppressor for colon cancer (Dong et al., 2019).

LINC00974 is a long intergenic non-protein coding gene. A study was done in 2021 (Shafabakhsh et al., 2021) proving that the LINC00974 gene is an RNA non-coding gene that significantly participates in the initiation, progression, and treatment resistance to Gastrointestinal (GI) cancers that generate colon cancer.

SPIB gene is a transcription factor of the E-twenty-six (ETS) gene family. SPIB gene also contributes to the invasiveness and metastasis of colon cancer (B.-Q. Li et al., 2012).

It was also discovered that CA7 gene (Carbonic anhydrase 7) expression in the RNA and protein is downregulated in colon cancer. The downregulated expression means it significantly reduces the differentiation of the disease and shortens patients' disease-specific survival.

The models have effectively detected the most correlated genes to a specific type of cancer from the biological analysis done on the common colon cancer genes.

The statistical analysis of the most potential biomarkers detected by the proposed models is shown in Figure (4.13), Figure (4.14), Figure (4.15), and Figure (4.16) for COAD, LUAD, BRCA, and COMBINED datasets, respectively. The selected top number of common genes of each model and the common genes between the models is illustrated. The numbers of the common genes were also described in Table (4.9). The figures illustrate the number of genes selected from each dataset. As shown in Figure (4.13), the number of common genes between MI-SVM and RF models is 9, RFE-SVM and RF are also 9, whereas the common genes between MI-SVM and RFE-SVM are 14. As a result, the number of common genes between all the models in the COAD dataset is five.

Similarly, for lung cancer in Figure (4.14), the number of common genes between MI-SVM and RF models is 2, RFE-SVM and RF are also 2, whereas the common genes between MI-SVM and RFE-SVM are 11. As a result, the number of common genes between all the models in the LUAD dataset is one. On the other hand, two common genes are obtained from the BRCA dataset, while 52 common genes are obtained from the COMBINED dataset, contributing to one or more cancer types.

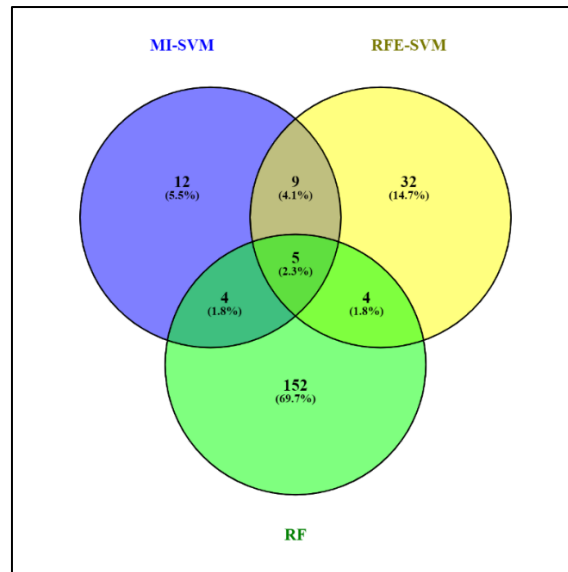


Figure (4.13): Venn Diagram of COAD Colon Cancer most Common Genes

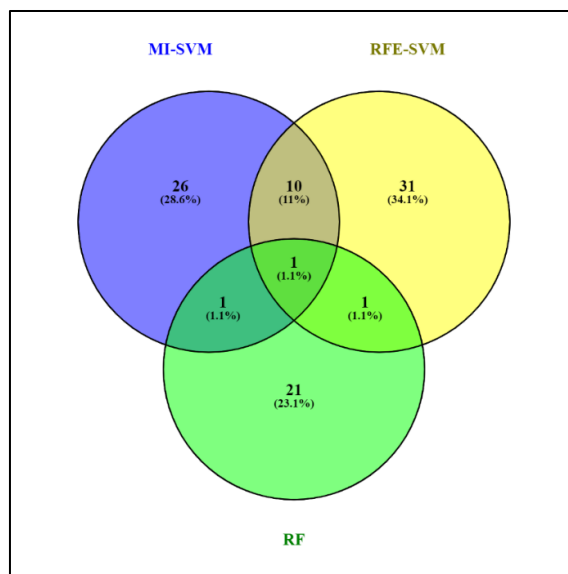


Figure (4.14): Venn Diagram of LUAD Lung Cancer most Common Genes

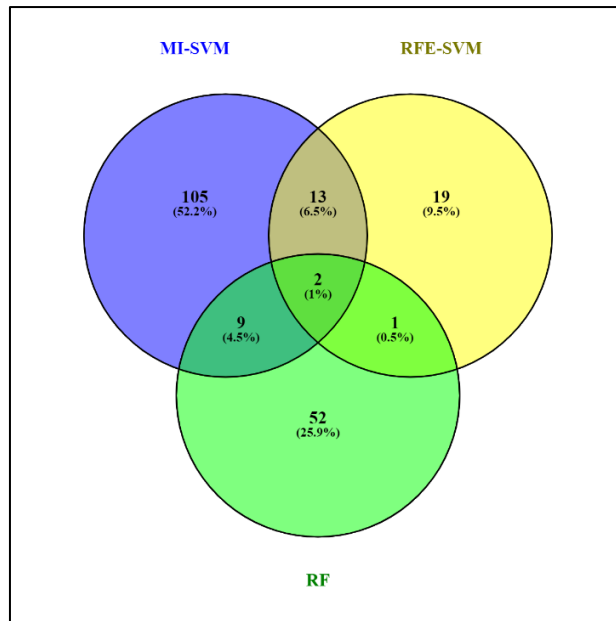


Figure (4.15): Venn Diagram of BRCA Breast Cancer most Common Genes

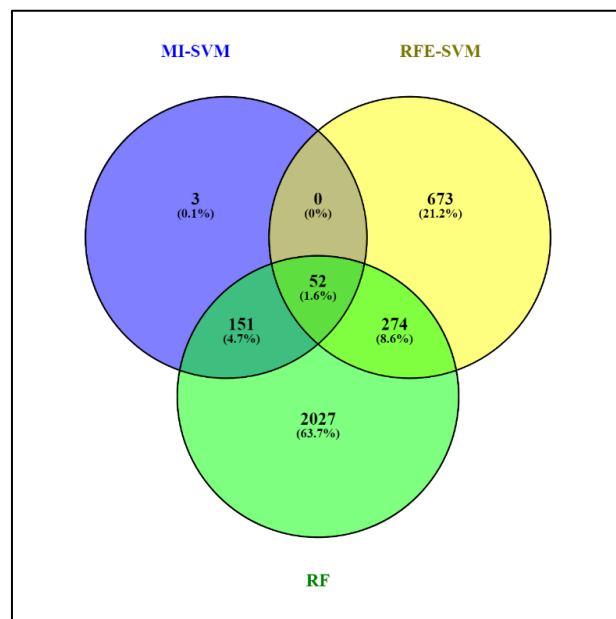


Figure (4.16): Venn Diagram of Combined Cancers Most Common Genes

4.4. Summary

The proposed models' experimental results emphasized the models' efficiency in predicting and classifying several types of cancers. The results also affirm the effectiveness of the models in selecting the most potential biomarkers; that may contribute to the development of cancer by analyzing the sequenced gene expression. The biological analysis study that was made on the selected genes assures the efficiency of the proposed models.

Chapter 5

Conclusion and Recommendations

5.1. Overview

5.2. Conclusion

5.3. Recommendation

Chapter 5: Conclusion and Recommendations

5.1. Overview

Chapter five is the concluding chapter of this thesis study. The chapter contains an overall summary of the thesis study. Contributions of the study are presented, and the essential findings of the proposed models are highlighted. It includes the conclusion and recommendations.

5.2. Conclusion

Cancer is a dangerous disease threatening humans' life. Cancer occurs due to abnormal growth in cells, which leads to their transformation into cancerous cells that may spread to other parts of the body. Therefore, there is a need for more studies that help detect cancer in the early stages to avoid the significant impacts on the patient's health that may lead to death. The significant advances in the NGS for genomic sequencing technologies have prompted many studies to reveal the effect of mutations in DNA and RNA genes on the development of cancer cells.

The possibility of adopting NGS technology in computer science has opened the way for scientists to use this technology in applying their knowledge in different bioinformatics applications. It enables scientists to study the genome and to analyze the whole genome structure. The whole-genome sequence and gene expression help the scientists detect the mutations in the genome sequence and the extent to which these mutations affect certain types of cancer using Artificial Intelligence (AI) technology. Machine learning technology is part of AI, and a lot of progress and development has occurred in this field. It has been used in the health sector

field to help physicians solve many problems that require a great effort in the analysis and diagnosis.

In this thesis study, NGS sequencing technology using RNA-Seq gene expression and machine learning algorithms were used to aid in the early detection of cancer and predict the probability of cancer by analyzing the genome and looking for specific mutations in the genes that may contribute to the development of cancer. Due to the massive size of the RNA-Seq datasets, different feature selection techniques should be implemented to handle the high dimensionality issue. Different machine learning techniques have been proposed in this study, and three models were implemented using different feature selection and classification techniques. The proposed models are MI-SVM, RFE-SVM, and RF. The models were implemented using four datasets, COAD, LUAD, BRCA cancers, and the fourth dataset, combining the previous three datasets. In the MI-SVM model, the most critical features (genes) were selected, and the model was trained on these selected genes. The second model, RFE-SVM, reduces the features using RFE as wrapper feature selection technique and classifies cancer accordingly. The third proposed model is the RF model. RF works as an embedded feature selection and classification algorithm that can efficiently detect the most important and related features. As part of the research objectives, all the proposed models' results were combined in an ensemble model. The model is trained using the combined list of genes to classify the most common biomarkers that contribute to a specific type of cancer. The results were validated and compared with several recently conducted research in this field. The obtained results (ranging from 98-100%) by the proposed models outperformed the compared results. The obtained lists of most potential biomarkers were analyzed more through biological analysis to ensure that

the genes obtained have already been confirmed as participatory for a particular type of cancer. The lists of obtained genes are illustrated in (Appendices Section).

5.3. Recommendation

In this thesis study, RNA-Seq gene expression was utilized using different machine learning algorithms and techniques. Due to the limitation of computer resources, deep learning techniques were not considered. Hence, different deep learning methods may be tested in the future.

Different input parameter values may be tested in the future for better results. Moreover, different feature selection and classification techniques may be used as combination hybrid or ensemble models to handle the high dimensionality and class imbalance issues.

According to the literature, several studies used the same RNA-Seq gene expression technique in predicting cancer survival opportunities. Hence, as future work, the proposed models may be used to study the survival opportunities. Several studies also used sequencing technology to detect the primary site of origin of cancer, and it can help in specialized medicine during patient treatment. It is an exciting research field that can be conducted in the future.

References

- Abraham, J., Heimberger, A. B., Marshall, J., Heath, E., Drabick, J., Helmstetter, A., ... Spetzler, D. (2021). Machine learning analysis using 77,044 genomic and transcriptomic profiles to accurately predict tumor type. <https://doi.org/10.1016/j.tranon.2021.101016>
- Adiwijaya, Wisesty, U. N., Lisnawati, E., Aditsania, A., & Kusumo, D. S. (2018). Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification. *Journal of Computer Science*, 14(11), 1521–1530. <https://doi.org/10.3844/jcssp.2018.1521.1530>
- Alanni, R., Hou, J., Azzawi, H., & Xiang, Y. (2019). A novel gene selection algorithm for cancer classification using microarray datasets. *BMC Medical Genomics*, 12(1), 1–12. <https://doi.org/10.1186/s12920-018-0447-6>
- Aroef, C., Yuda, R. P., Rustam, Z., & Pandelaki, J. (2019). Multinomial Logistic Regression and Support Vector Machine for Osteoarthritis Classification. *Journal of Physics: Conference Series*, 1417(1). <https://doi.org/10.1088/1742-6596/1417/1/012012>
- Arowol, M. O., Abdulsalam, S. O., Isiaka, R. M., & Gbolagade, K. A. (2017). A Hybrid Dimensionality Reduction Model for Classification of Microarray Dataset. *International Journal of Information*

Technology and Computer Science, 9(11), 57–63.

<https://doi.org/10.5815/ijitcs.2017.11.06>

Aswathy, M. A., & Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: Present status and future possibilities.

Informatics in Medicine Unlocked, 8, 74–79.

<https://doi.org/10.1016/j.imu.2016.11.001>

Ayyad, S. M., Saleh, A. I., & Labib, L. M. (2019). Gene expression cancer classification using modified K-Nearest Neighbors technique.

BioSystems, 176(December 2018), 41–51.

<https://doi.org/10.1016/j.biosystems.2018.12.009>

Ballhausen, A., Przybilla, M. J., Jendrusch, M., Haupt, S., Pfaffendorf, E.,

Seidler, F., ... Kloor, M. (2020). The shared frameshift mutation

landscape of microsatellite-unstable cancers suggests immunoediting

during tumor evolution. *Nature Communications*, 11(1), 4740.

<https://doi.org/10.1038/s41467-020-18514-5>

Bock, H.-H. (2007). *Clustering Methods: A History of k-Means*

Algorithms. (1957), 161–172. [https://doi.org/10.1007/978-3-540-](https://doi.org/10.1007/978-3-540-73560-1_15)

[73560-1_15](https://doi.org/10.1007/978-3-540-73560-1_15)

Brendan P. Hodkinson Elizabeth A. Grice. (2015). Next-Generation

Sequencing: A Review of Technologies and Tools for Wound

Microbiome Research. *Advances in Wound Care*, 4(1).

Çelik, A., Yaman, H., Turan, S., Kara, A., Kara, F., Zhu, B., ... Dutta, D.

- (2018). An Evidence-Based Approach. In *Journal of Materials Processing Technology* (Vol. 1). Retrieved from <http://dx.doi.org/10.1016/j.cirp.2016.06.001><http://dx.doi.org/10.1016/j.powtec.2016.12.055><https://doi.org/10.1016/j.ijfatigue.2019.02.006><https://doi.org/10.1016/j.matlet.2019.04.024><https://doi.org/10.1016/j.matlet.2019.127252><http://dx.doi.org/10.1016/j.matlet.2019.127252>
- Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408, 189–215. <https://doi.org/10.1016/j.neucom.2019.10.118>
- Chang, C. C., & Chen, S. H. (2019). Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Breast Cancer Survivors. *Frontiers in Genetics*, 10. <https://doi.org/10.3389/fgene.2019.00848>
- Chen, Y., Huang, J., He, X., Gao, Y., Mahara, G., Lin, Z., & Zhang, J. (2019). A novel approach to determine two optimal cut-points of a continuous predictor with a U-shaped relationship to hazard ratio in survival data: Simulation and application. *BMC Medical Research Methodology*, 19(1), 1–12. <https://doi.org/10.1186/s12874-019-0738-4>
- Ching, T., Zhu, X., & Garmire, L. X. (2018). *Cox-nnet : An artificial neural network method for prognosis prediction of high-throughput*

omics data. 1–18.

Colombo, P. E., Milanezi, F., Weigelt, B., & Reis-Filho, J. S. (2011).

Microarrays in the 2010s: The contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Research*, 13(3).

<https://doi.org/10.1186/bcr2890>

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., ...

Velculescu, V. E. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761), 385–389.

<https://doi.org/10.1038/s41586-019-1272-6>

Cueto-lópez, N., García-ordás, M. T., Dávila-batista, V., Moreno, V.,

Aragonés, N., & Alaiz-rodríguez, R. (2019). A comparative study on feature selection for a risk prediction model for colorectal cancer.

Computer Methods and Programs in Biomedicine, 177, 219–229.

<https://doi.org/10.1016/j.cmpb.2019.06.001>

Cunningham, P., & Delany, S. J. (2007). K -Nearest Neighbour

Classifiers. *Multiple Classifier Systems*, (May), 1–17.

Dong, Y., Zhang, Y., Kang, W., Wang, G., Chen, H., Higashimori, A., ...

Yu, J. (2019). VSTM2A suppresses colorectal cancer and antagonizes Wnt signaling receptor LRP6. *Theranostics*, 9, 22.

<https://doi.org/10.7150/thno.34989>

Evgeniou, T., & Pontil, M. (2001). Support vector machines: Theory and

applications. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2049 LNAI(May), 249–257.

https://doi.org/10.1007/3-540-44673-7_12

Expression of VSTM2A in Colorectal Cancer - The Human Protein Atlas.

(2021). Retrieved March 10, 2021, from

<https://www.proteinatlas.org/ENSG00000170419->

VSTM2A/pathology/colorectal+cancer/READ

Fawagreh, K., Medhat Gaber, M., Elyan, E., & Gaber, M. M. (2014).

Random forests: from early developments to recent advancements.

Systems Science & Control Engineering: An Open Access Journal,

2(1), 602–609. <https://doi.org/10.1080/21642583.2014.956265>

García-Pedrajas, N., de Haro-García, A., & Perez-Rodríguez, J. (2014). A

scalable memetic algorithm for simultaneous instance and feature

selection. *Evolutionary Computation*, 22(1), 1–45.

https://doi.org/10.1162/EVCO_a_00102

GDC Portal. (2021). Retrieved from <https://portal.gdc.cancer.gov/>

Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., & Maulik, U. (2019).

Recursive Memetic Algorithm for gene selection in microarray data.

Expert Systems with Applications, 116, 172–185.

<https://doi.org/10.1016/j.eswa.2018.06.057>

Goossens, N., Nakagawa, S., Sun, X., & Hoshida, Y. (2015). Cancer

- biomarker discovery and validation. *Translational Cancer Research*, 4(3). <https://doi.org/10.3978/j.issn.2218-676X.2015.06.04>
- Guo, Y., Liu, S., Li, Z., & Shang, X. (2018). BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC Bioinformatics*, 19(Suppl 5), 1–13. <https://doi.org/10.1186/s12859-018-2095-4>
- Hancer, E. (2018). Differential evolution for feature selection : a fuzzy wrapper – filter approach. *Soft Computing*, (Jolliffe 2014). <https://doi.org/10.1007/s00500-018-3545-7>
- Hooshmand, A. (2020). Accurate diagnosis of Cancer by machine learning classification of the whole genome sequencing data. *ArXiv*, 1–30.
- Huljanah, M., Rustam, Z., Utama, S., & Siswantining, T. (2019). Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. *IOP Conference Series: Materials Science and Engineering*, 546(5). <https://doi.org/10.1088/1757-899X/546/5/052031>
- Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E., & Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nature Microbiology*, 5(10), 1202–1206. <https://doi.org/10.1038/s41564-020-0774-1>
- Jain, D., & Singh, V. (2018). An Efficient Hybrid Feature Selection

- model for Dimensionality Reduction. *Procedia Computer Science*, 132(Iccids), 333–341. <https://doi.org/10.1016/j.procs.2018.05.188>
- Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Al-Shahrour, F., ... Stein, L. D. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, 11(1), 1–12. <https://doi.org/10.1038/s41467-019-13825-8>
- Jing, B., Zhang, T., Wang, Z., Jin, Y., Liu, K., Qiu, W., & Li, C. (2019). Artificial Intelligence In Medicine A deep survival analysis method based on ranking. *Artificial Intelligence In Medicine*, 98(September 2018), 1–9. <https://doi.org/10.1016/j.artmed.2019.06.001>
- Jolfaei, N. A., Jolfaei, N. A., Hekmatifar, M., Piranfar, A., Toghraie, D., Sabetvand, R., & Rostami, S. (2020). Investigation of thermal properties of DNA structure with precise atomic arrangement via equilibrium and non-equilibrium molecular dynamics approaches. *Computer Methods and Programs in Biomedicine*, 185, 105169. <https://doi.org/10.1016/j.cmpb.2019.105169>
- Katoch, S., Chauhan, S. S., & Kumar, V. (2020). A review on genetic algorithm: past, present, and future. In *Multimedia Tools and Applications*. Multimedia Tools and Applications. <https://doi.org/10.1007/s11042-020-10139-6>
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., & Jiang, T. (2018).

DeepSurv : personalized treatment recommender system using a Cox proportional hazards deep neural network. 1–12.

- Kim, A. A., Rachid Zaim, S., & Subbian, V. (2020). Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data. *International Journal of Medical Informatics*, 141(April), 104148.
<https://doi.org/10.1016/j.ijmedinf.2020.104148>
- Kucharska, E. (2019). Deep Neural Network Algorithm Feedback Model with Behavioral Intelligence and Forecast Accuracy. *Mdpi*, 1–15.
- Lee, E. P. F., Lee, E. P. F., Lozeille, J., Soldán, P., Daire, S. E., Dyke, J. M., & Wright, T. G. (2001). An empirical study of the naive Bayes classifier. *Physical Chemistry Chemical Physics*, 3(22), 4863–4869.
<https://doi.org/10.1039/b104835j>
- Li, B.-Q., Huang, T., Liu, L., Cai, Y.-D., Chou, K.-C., Lee Ho, P., & Butantan, I. (2012). Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network. *LoS ONE*. <https://doi.org/10.1371/journal.pone.0033393>
- Li, H., Tian, S., Li, Y., Fang, Q., Tan, R., Pan, Y., ... Gao, X. (2020). Modern deep learning in bioinformatics. *Journal of Molecular Cell Biology*, 12(11), Pages 823–827.
<https://doi.org/10.1093/jmcb/mjaa030>
- Liao, W., Liu, F., Zhang, H., Liao, W., Xu, N., Xie, W., & Zhang, Y.

- (2020). *Upregulation of GpncA is associated with poor prognosis through enhancement of tumor growth via regulating GSK3B*.
<https://doi.org/10.1038/s41598-020-58729-6>
- Lin, X., Li, C., Zhang, Y., Su, B., & Wei, H. (2018). molecules Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics. *Molecules*, 23(1), 52.
<https://doi.org/10.3390/molecules23010052>
- Ling, C. X., & Sheng, V. S. (2010). Class Imbalance Problem. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (p. 171). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_110
- Liu, Y., Salamov, A., Gao, W., Rodin, A. S., Wang, X., Branciamore, S., ... Ding, S. (2020). *New Analysis Framework Incorporating Mixed Mutual Information and Scalable Bayesian Networks for Multimodal High Dimensional Genomic and Epigenomic Cancer Data*.
<https://doi.org/10.3389/fgene.2020.00648>
- Liu, Z., Tang, D., Cai, Y., Wang, R., & Chen, F. (2017). A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. *Neurocomputing*, 266, 641–650.
<https://doi.org/10.1016/j.neucom.2017.05.066>
- Love, S., & McLean, C. (2011). Overview and recent advances in neuropathology. *Pathology*, 43(2), 87.

<https://doi.org/10.1097/PAT.0b013e3283429f6c>

Ma, J., & Gao, X. (2020). Knowledge-Based Systems A filter-based feature construction and feature selection approach for classification using Genetic Programming. *Knowledge-Based Systems*, 196(0950–7051), 105806. <https://doi.org/10.1016/j.knosys.2020.105806>

Mardis, E. R., & Wilson, R. K. (2009). Cancer genome sequencing: a review. *Human Molecular Genetics*, 18(R2), R163–R168. <https://doi.org/10.1093/hmg/ddp396>

Michal Janitz. (2011). *Next Generation Genome Sequencing: Towards Personalized Medicine*.

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458), 454–455. [https://doi.org/10.1016/s0140-6736\(05\)70272-5](https://doi.org/10.1016/s0140-6736(05)70272-5)

Mohan, P., & Paramasivam, I. (2020). *Feature reduction using SVM-RFE technique to detect autism spectrum disorder. 1, 3*. <https://doi.org/10.1007/s12065-020-00498-2>

Nannini, M., Pantaleo, M. A., Maleddu, A., Astolfi, A., Formica, S., & Biasco, G. (2009). Gene expression profiling in colorectal cancer using microarray technologies: Results and perspectives. *Cancer Treatment Reviews*, 35(3), 201–209. <https://doi.org/10.1016/j.ctrv.2008.10.006>

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial.

Frontiers in Neurorobotics, 7(DEC).

<https://doi.org/10.3389/fnbot.2013.00021>

P, S. S., H, K. A., & Ghaznavi, H. (2020). Predicting Lung Cancer Patients' Survival Time via Logistic Regression-based Models in a Quantitative Radiomic Framework. *J Biomed Phys Eng*, 10(4).

<https://doi.org/10.31661/jbpe.v0i0.1027>

Pavithra, D. (2017). *Feature Selection and Classification in gene expression cancer data*.

Penson, A., Camacho, N., Zheng, Y., Varghese, A. M., Al-Ahmadie, H., Razavi, P., ... Berger, M. F. (2020). Development of Genome-Derived Tumor Type Prediction to Inform Clinical Cancer Care. *JAMA Oncology*, 6(1), 84–91.

<https://doi.org/10.1001/jamaoncol.2019.3985>

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, (March), 61–74.

Pothuganti, K. (2020). OVERVIEW ON PRINCIPAL COMPONENT ANALYSIS ALGORITHM IN MACHINE LEARNING. *International Research Journal of Modernization in Engineering Technology and Science*, 02(10).

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor,

- T. R., ... Gu, Y. (2012). A tale of three NGS sequencing platforms. *BMC Genomics*, 13(341), 13.
- Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(1), 1–14. <https://doi.org/10.1186/s12859-016-1423-9>
- Ramírez-Hernández, J. A., & Fernandez, E. (2007). Enhanced recursive feature elimination. *Proceedings - 6th International Conference on Machine Learning and Applications, ICMLA*, (June), 330–335. <https://doi.org/10.1109/ICMLA.2007.35>
- Rasmussen, C. E., & Kuss, M. (2004). Gaussian processes in reinforcement learning. *Advances in Neural Information Processing Systems*, (1).
- Rassy, E., Assi, T., & Pavlidis, N. (2020). Exploring the biological hallmarks of cancer of unknown primary: where do we stand today? *British Journal of Cancer*, 122. <https://doi.org/10.1038/s41416-019-0723-z>
- Ross, B. C. (2014). *Mutual Information between Discrete and Continuous Data Sets*. <https://doi.org/10.1371/journal.pone.0087357>
- Sahran, S., Albashish, D., Abdullah, A., Shukor, N. A., & Pauzi, S. H. (2018). Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading. *Artificial Intelligence In*

Medicine. <https://doi.org/10.1016/j.artmed.2018.04.002>

Sapna, S. (2012). Backpropagation Learning Algorithm Based on

Levenberg Marquardt Algorithm. *Airccj*, 393–398.

<https://doi.org/10.5121/csit.2012.2438>

Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A Nested Genetic

Algorithm for feature selection in high-dimensional cancer

Microarray datasets. *Expert Systems with Applications*, 121, 233–

243. <https://doi.org/10.1016/j.eswa.2018.12.022>

Schneider, A., Hommel, G., & Blettner, M. (2010). *Linear Regression*

Analysis. 107(44), 776–782.

<https://doi.org/10.3238/arztebl.2010.0776>

Shafabakhsh, R., Arianfar, F., Vosough, M., Mirzaei, H. R., Mahjoubin,

M., Khanbabaei, H., ... Mirzaei, H. (2021). Autophagy and

gastrointestinal cancers: the behind the scenes role of long non-

coding RNAs in initiation, progression, and treatment resistance.

Cancer Gene Therapy.

Shalaby, T., Fiaschetti, G., Baumgartner, M., & Grotzer, M. A. (2014).

MicroRNA Signatures as Biomarkers and Therapeutic Target for

CNS Embryonal Tumors: The Pros and the Cons. *OPEN ACCESS*

Int. J. Mol. Sci, 15, 15. <https://doi.org/10.3390/ijms151121554>

Sharma, A., & Rani, R. (2021). A Systematic Review of Applications of

Machine Learning in Cancer Prediction and Diagnosis. *Archives of*

Computational Methods in Engineering.

<https://doi.org/10.1007/s11831-021-09556-z>

Sharmin, S., Shoyaib, M., Ali, A. A., Khan, M. A. H., & Chae, O. (2019).

Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition*, 91, 162–174.

<https://doi.org/10.1016/j.patcog.2019.02.016>

Sherafatian, M. (2018). Tree-based machine learning algorithms

identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene*, 677, 111–118.

<https://doi.org/10.1016/j.gene.2018.07.057>

Sherafatian, M., & Arjmand, F. (2019). Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data. *Oncology Letters*, 18(2), 2125–2131.

<https://doi.org/10.3892/ol.2019.10462>

Shon, H. S., Batbaatar, E., Kim, K. O., Cha, E. J., & Kim, K. A. (2020).

Classification of kidney cancer data using cost-sensitive hybrid deep learning approach. *Symmetry*, (1).

<https://doi.org/10.3390/sym12010154>

Şirin, N., Elmas, L., Seçme, M., & Dodurga, Y. (2020). Investigation of possible effects of apigenin, sorafenib and combined applications on apoptosis and cell cycle in hepatocellular cancer cells. *Gene*, 737,

144428. <https://doi.org/10.1016/j.gene.2020.144428>

- Soh, K. P., Szczurek, E., Sakoparnig, T., & Beerenwinkel, N. (2017). Predicting cancer type from tumour DNA signatures. *Genome Medicine*, 9(1), 1–12. <https://doi.org/10.1186/s13073-017-0493-2>
- Sonneveld, S., Verhagen, B. M. P., & Tanenbaum, M. E. (2020). Heterogeneity in mRNA Translation. *Trends in Cell Biology*, 30(8), 606–618. <https://doi.org/10.1016/j.tcb.2020.04.008>
- Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9, 1–10. <https://doi.org/10.1186/1471-2105-9-319>
- Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in Genetics*, 10(MAR), 1–13. <https://doi.org/10.3389/fgene.2019.00256>
- Thakur, V. (2015). Why Nature Preferred DNA over RNA? - Science Samhita. Retrieved from <https://sciencesamhita.com/dna-as-the-genetic-material/>
- Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2), 145. <https://doi.org/10.1504/ijapr.2016.079050>
- Tissue expression of BEST4 - Staining in colon - The Human Protein Atlas. (2021). Retrieved from

<https://www.proteinatlas.org/ENSG00000142959->

BEST4/tissue/colon

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J.

H. (2018). Relief-based feature selection: Introduction and review.

Journal of Biomedical Informatics, 85(July), 189–203.

<https://doi.org/10.1016/j.jbi.2018.07.014>

Venkataramana, L., Gracia, S., Rajavel, J., & Dodda, R. (2019).

Improving classification accuracy of cancer types using parallel

hybrid feature selection on microarray gene expression data. *Genes &*

Genomics, (0123456789). <https://doi.org/10.1007/s13258-019->

00859-x

Von Felden, J., Craig, A. J., Garcia-Lezana, T., Labgaa, • Ismail, Haber,

P. K., Delia D'avola, •, ... Villanueva, • Augusto. (2021). Mutations

in circulating tumor DNA predict primary resistance to systemic

therapies in advanced hepatocellular carcinoma. *Oncogene*, 40, 140–

151. <https://doi.org/10.1038/s41388-020-01519-1>

Wah, Y. B., Ibrahim, N., & Hamid, H. A. (2018). Feature selection

methods : Case of filter and wrapper approaches for maximising

classification accuracy. *SCIENCE & TECHNOLOGY*, 26(1), 329–

340.

Wan, N., Weinberg, D., Liu, T. Y., Niehaus, K., Delubac, D., Kannan, A.,

... Haque, I. S. (2018). Machine learning enables detection of early-

- stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BioRxiv*, 1–22. <https://doi.org/10.1101/478065>
- Wan, X. (2019). To cite this article: Xing Wan. *J. Phys.* <https://doi.org/10.1088/1742-6596/1213/3/032021>
- Wang, J., Xie, X., Shi, J., He, W., Chen, Q., Chen, L., ... Zhou, T. (2020). Denoising Autoencoder, a Deep Learning Algorithm, Aids the Identification of a Novel Molecular Signature of Lung Adenocarcinoma. *Genomics, Proteomics & Bioinformatics*. <https://doi.org/10.1016/j.gpb.2019.02.003>
- Wang, Q., Zhou, Y., Zhang, W., Tang, Z., & Chen, X. (2020). Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis. *Expert Systems with Applications*, 152, 113334. <https://doi.org/10.1016/j.eswa.2020.113334>
- Wishart, D. S. (2015). EBioMedicine Is Cancer a Genetic Disease or a Metabolic Disease ? *EBIOM*, 2(6), 478–479. <https://doi.org/10.1016/j.ebiom.2015.05.022>
- Wu, Y. (2020). *Feature Selection on High Dimensional Histogram Data to Improve Vehicle Components' LifeComponents' Life Length Prediction*. Retrieved from <http://www.teknat.uu.se/student>
- Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018a). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153, 1–9.

<https://doi.org/10.1016/j.cmpb.2017.09.005>

Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018b). A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Computer Methods and Programs in Biomedicine*, 166, 99–105.

<https://doi.org/10.1016/j.cmpb.2018.10.004>

Xu, L., Wang, R., Ziegelbauer, J., Wu, W. W., Shen, R.-F., Juhl, H., ... Rosenberg, A. S. (2017). *Transcriptome analysis of human colorectal cancer biopsies reveals extensive expression correlations among genes related to cell proliferation, lipid metabolism, immune response and collagen catabolism*. Retrieved from www.impactjournals.com/oncotarget

Yerukala Sathipati, S., & Ho, S. Y. (2018). Identifying a miRNA signature for predicting the stage of breast cancer. *Scientific Reports*, 8(1), 1–11. <https://doi.org/10.1038/s41598-018-34604-3>

Zang, C., Wang, T., Deng, K., Li, B., Hu, E., Qin, Q., ... Shirley Liu, X. (2016). High-dimensional genomic data bias correction and data integration using MANCIE. *Nature Communications*, 7(11305). <https://doi.org/10.1038/ncomms11305>

Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Medical and Biological Engineering and*

Computing, 57(4), 901–912. <https://doi.org/10.1007/s11517-018-1930-0>

Zhu, Y., Lai, M., Pan, F., Chen, T., Sun, X., Li, K., ... Försti, A. (2019). Prognosis Prediction of Colorectal Cancer Using Gene Expression Profiles. *Frontiers in Oncology*, 1, 252.
<https://doi.org/10.3389/fonc.2019.00252>

Zong, W., Huang, G. Bin, & Chen, Y. (2013). Weighted extreme learning machine for imbalance learning. *Neurocomputing*, 101, 229–242.
<https://doi.org/10.1016/j.neucom.2012.08.010>

Zou, Q., Wei, L., Basu, M., Cai, L., Liao, B., Yang, J., ... Chen, Y. (2019). Improved Pre-miRNAs Identification Through Mutual Information of Pre-miRNA Sequences and Structures. *Frontiers in Genetics* | *Www.Frontiersin.Org*, 10, 119.
<https://doi.org/10.3389/fgene.2019.00119>

Appendices

Appendix (1)

Colon Cancer Potential Biomarkers (Genes)

Colon Genes selected by MI-SVC Model		
ENSG00000104267	CA2	
ENSG00000197273	GUCA2A	
ENSG00000184544	DHRS7C	
ENSG00000123560	PLP1	
ENSG00000187699	C2orf88	
ENSG00000163106	HPGDS	
ENSG00000108231	LGI1	
ENSG00000167767	KRT80	
ENSG00000141338	ABCA8	
ENSG00000115263	GCG	
ENSG00000163412	EIF4E3	
ENSG00000164283	ESM1	
ENSG00000152785	BMP3	
ENSG00000168546	GFRA2	
ENSG00000226629	LINC00974	
ENSG00000108381	ASPA	
ENSG00000170419	VSTM2A	
ENSG00000182271	TMIGD1	
ENSG00000186198	SLC51B	
ENSG00000164303	ENPP6	
ENSG00000269404	SPIB	
ENSG00000135549	PKIB	
ENSG00000160712	IL6R	
ENSG00000111846	GCNT2	
ENSG00000183034	OTOP2	
ENSG00000168748	CA7	
ENSG00000168079	SCARA5	
ENSG00000065325	GLP2R	
ENSG00000062038	CDH3	
ENSG00000142959	BEST4	

Colon Genes selected by RFE-SVC Model			
Gene ID	Gene Name	Gene ID	Gene Name
ENSG00000079689	SCGN	ENSG00000242752	NMTRQ-TTG12-1
ENSG00000080493	SLC4A4	ENSG00000253286	AC135166.1
ENSG00000101842	VSIG1	ENSG00000253881	AC105233.1
ENSG00000102683	SGCG	ENSG00000260897	DNM1P30
ENSG00000104267	CA2	ENSG00000265266	
ENSG00000111404	RERGL	ENSG00000266176	
ENSG00000111846	GCNT2	ENSG00000269404	SPIB
ENSG00000125144	MT1G	ENSG00000279756	AC012317.2
ENSG00000131096	PYY		
ENSG00000133742	CA1		
ENSG00000134028	ADAMDEC1		
ENSG00000137225	CAPN11		
ENSG00000141338	ABCA8		
ENSG00000142698	C1orf94		
ENSG00000142959	BEST4		
ENSG00000143171	RXRG		
ENSG00000152785	BMP3		
ENSG00000159197	KCNE2		
ENSG00000163815	CLEC3B		
ENSG00000166869	CHP2		
ENSG00000167434	CA4		
ENSG00000168079	SCARA5		
ENSG00000168748	CA7		
ENSG00000170419	VSTM2A		
ENSG00000182271	TMIGD1		
ENSG00000183034	OTOP2		
ENSG00000184544	DHRS7C		
ENSG00000186198	SLC51B		
ENSG00000187783	TMEM72		
ENSG00000189129	PLAC9		
ENSG00000196666	FAM180B		
ENSG00000198788	MUC2		
ENSG00000203818			
ENSG00000204936	CD177		
ENSG00000205364	MT1M		
ENSG00000206615	RNU6-338P		
ENSG00000214100	PLAC9P1		
ENSG00000226087	AC073283.2		
ENSG00000226629	LINC00974		
ENSG00000228776	AC114492.1		
ENSG00000229188	AL050349.1		
ENSG00000236627	AL590302.2		

Colon Genes selected by RF Model			
Gene ID	Gene Name	Gene ID	Gene Name
ENSG00000142959	BEST4	ENSG00000173452	TMEM196
ENSG00000100604	CHGA	ENSG00000090006	LTBP4
ENSG00000170419	VSTM2A	ENSG00000261513	AC009097.4
ENSG00000226629	LINC00974	ENSG00000267257	AC105105.1
ENSG00000104435	STMN2	ENSG00000134962	KLB
ENSG00000102683	SGCG	ENSG00000169129	AFAP1L2
ENSG00000166869	CHP2	ENSG00000258247	
ENSG00000168447	SCNN1B	ENSG00000163683	SMIM14
ENSG00000155850	SLC26A2	ENSG00000165935	SMCO2
ENSG00000109182	CWH43	ENSG00000187848	P2RX2
ENSG00000110975	SYT10	ENSG00000198961	PJA2
ENSG00000040199	PHLPP2	ENSG00000154080	CHST9
ENSG00000132915	PDE6A	ENSG00000131019	ULBP3
ENSG00000091138	SLC26A3	ENSG00000160712	IL6R
ENSG00000182938	OTOP3	ENSG00000124205	EDN3
ENSG00000145708	CRHBP	ENSG00000189067	LITAF
ENSG00000138722	MMRN1	ENSG00000253250	C8orf88
ENSG00000134201	GSTM5	ENSG00000258021	AC025259.2
ENSG00000129167	TPH1	ENSG00000022267	FHL1
ENSG00000185269	NOTUM	ENSG00000271590	AC108463.3
ENSG00000186462	NAP1L2	ENSG00000132471	WBP2
ENSG00000165072	MAMDC2	ENSG00000229274	LINC02829
ENSG00000113396	SLC27A6	ENSG00000186564	FOXD2
ENSG00000137727	ARHGAP20	ENSG00000108700	CCL8
ENSG00000111802	TDP2	ENSG00000228203	GRASLND
ENSG00000160408	ST6GALNAC6	ENSG00000164283	ESM1
ENSG00000116016	EPAS1	ENSG00000172005	MAL
ENSG00000228401	AL590226.1	ENSG00000151632	AKR1C2
ENSG00000145287	PLAC8	ENSG00000104332	SFRP1
ENSG00000271447	MMP28	ENSG00000186007	LEMD1
ENSG00000168309	FAM107A	ENSG00000149218	ENDOD1
ENSG00000164342	TLR3	ENSG00000232995	RGS5
ENSG00000119138	KLF9	ENSG00000163106	HPGDS
ENSG00000008382	MPND	ENSG00000225698	IGHV3-72
ENSG00000256632	AC135586.2	ENSG00000246308	LOC101928053
ENSG00000108187	PBLD	ENSG00000108924	HLF
ENSG00000141434	MEP1B	ENSG00000251022	THAP9-AS1
ENSG00000115762	PLEKHB2	ENSG00000169330	MINAR1
ENSG00000115665	SLC5A7	ENSG00000174992	ZG16
ENSG00000272143	FGF14-AS2	ENSG00000079689	SCGN
ENSG00000260173	AC020779.2	ENSG00000269927	AC004817.3
ENSG00000138744	NAAA	ENSG00000256008	AC125616.1
ENSG00000250722	SELENOP	ENSG00000173588	CEP83

ENSG00000198121	LPAR1	ENSG00000105865	DUS4L
ENSG00000155918	RAET1L	ENSG00000181450	ZNF678
ENSG00000130226	DPP6	ENSG00000111249	CUX2
ENSG00000269404	SPIB	ENSG00000247746	USP51
ENSG00000224968	LINC01645	ENSG00000178395	CCDC185
ENSG00000189037	DUSP21	ENSG00000115295	CLIP4
ENSG00000160124	MIX23	ENSG00000265243	IGLJCOR18
ENSG00000108602	ALDH3A1	ENSG00000248866	USP46-DT
ENSG00000004799	PDK4	ENSG00000224003	YES1P1
ENSG00000166415	WDR72	ENSG00000219682	AL133268.1
ENSG00000145526	CDH18	ENSG00000211958	IGHV3-38
ENSG00000168748	CA7	ENSG00000167641	PPP1R14A
ENSG00000124302	CHST8	ENSG00000140688	RUSF1
ENSG00000036672	USP2	ENSG00000138193	PLCE1
ENSG00000268945	AC010422.2	ENSG00000102048	ASB9
ENSG00000257660	ADCY6-DT	ENSG00000070748	CHAT
ENSG00000149294	NCAM1	ENSG00000254645	AC087379.1
ENSG00000135437	RDH5	ENSG00000234690	EPCAM-DT
ENSG00000169925	BRD3	ENSG00000230384	RPSAP25
ENSG00000277653	AL354709.2	ENSG00000215482	CALM2P3
ENSG00000274667	AC090517.2	ENSG00000130595	TNNT3
ENSG00000187556	NANOS3	ENSG00000273026	AL358472.3
ENSG00000164087	POC1A	ENSG00000227066	Z98257.1
ENSG00000137098	SPAG8	ENSG00000184451	CCR10
ENSG00000012504	NR1H4	ENSG00000170925	TEX13B
ENSG00000280309	AC023300.3	ENSG00000102882	MAPK3
ENSG00000251186	AC105345.2	ENSG00000074842	MYDGF
ENSG00000170915	PAQR8	ENSG00000241933	DENND6A-DT
ENSG00000137474	MYO7A	ENSG00000170345	FOS
ENSG00000163347	CLDN1	ENSG00000253549	CA3-AS1
ENSG00000138363	ATIC	ENSG00000184156	KCNQ3
ENSG00000118733	OLFM3	ENSG00000176399	DMRTA1
ENSG00000247556	OIP5-AS1	ENSG00000172733	PURG
ENSG00000204118	NAP1L6P	ENSG00000236915	CLCA4-AS1
ENSG00000146926	ASB10	ENSG00000168546	GFRA2
ENSG00000116922	C1orf109		
ENSG00000100003	SEC14L2		
ENSG00000099817	POLR2E		
ENSG00000241808	RPS15AP34		
ENSG00000215177	IGLV8OR8-1		
ENSG00000133742	CA1		
ENSG00000278535	DHRS11		
ENSG00000271443	AC011247.2		
ENSG00000210144	MT-TY		

Colon Common Selected Genes between MI-RFE Models	
Gene ID	Gene Name
ENSG00000142959	BEST4
ENSG00000170419	VSTM2A
ENSG00000226629	LINC00974
ENSG00000104435	STMN2
ENSG00000186462	NAP1L2
ENSG00000137727	ARHGAP20
ENSG00000160712	IL6R
ENSG00000164283	ESM1
ENSG00000163106	HPGDS
ENSG00000079689	SCGN
ENSG00000269404	SPIB
ENSG00000004799	PDK4
ENSG00000168748	CA7
ENSG00000135437	RDH5
ENSG00000168546	GFRA2

Colon Common Selected Genes between RFE-RF Models	
Gene ID	Gene Name
ENSG00000142959	BEST4
ENSG00000170419	VSTM2A
ENSG00000226629	LINC00974
ENSG00000102683	SGCG
ENSG00000166869	CHP2
ENSG00000079689	SCGN
ENSG00000269404	SPIB
ENSG00000168748	CA7
ENSG00000133742	CA1

Colon Common Selected Genes between MI-RFE-RF Models	
Gene ID	Gene Name
ENSG00000142959	BEST4
ENSG00000170419	VSTM2A
ENSG00000226629	LINC00974
ENSG00000269404	SPIB
ENSG00000168748	CA7

Appendix (2)

Lung Cancer Potential Biomarkers (Genes)

Lung Genes selected by MI-SVC Model		
Gene ID	Gene Name	ENSG00000204305 AGER
ENSG00000261863	LINC01996	
ENSG00000011426	ANLN	
ENSG000000085999	RAD54L	
ENSG00000112782	CLIC5	
ENSG00000169607	CKAP2L	
ENSG00000169252	ADRB2	
ENSG00000168309	FAM107A	
ENSG00000121211	MND1	
ENSG00000129173	E2F8	
ENSG00000108576	SLC6A4	
ENSG00000128050	PAICS	
ENSG00000156970	BUB1B	
ENSG00000163815	CLEC3B	
ENSG00000024526	DEPDC1	
ENSG00000126787	DLGAP5	
ENSG00000224215	LOC105376453	
ENSG00000167588	GPD1	
ENSG00000204949	FAM83A-AS1	
ENSG00000091651	ORC6	
ENSG00000101280	ANGPT4	
ENSG00000066405	CLDN18	
ENSG00000138160	KIF11	
ENSG00000165480	SKA3	
ENSG00000197465	GYPE	
ENSG00000183856	IQGAP3	
ENSG00000123485	HJURP	
ENSG00000150625	GPM6A	
ENSG00000186871	ERCC6L	
ENSG00000196584	XRCC2	
ENSG00000066279	ASPM	
ENSG00000276043	UHRF1	
ENSG00000168484	SFTPC	
ENSG00000135604	STX11	
ENSG00000170323	FABP4	
ENSG00000158402	CDC25C	
ENSG00000131747	TOP2A	
ENSG00000234281	LANCL1-AS1	

Lung Genes selected by RFE-SVC Model		
Gene ID	Gene Name	ENSG00000263632
ENSG00000073754	CD5L	ENSG00000266321 MIR4293
ENSG00000101280	ANGPT4	
ENSG00000107742	SPOCK2	
ENSG00000108576	SLC6A4	
ENSG00000114854	TNNC1	
ENSG00000122679	RAMP3	
ENSG00000131355	ADGRE3	
ENSG00000132671	SSTR4	
ENSG00000135604	STX11	
ENSG00000140600	SH3GL3	
ENSG00000142748	FCN3	
ENSG00000148677	ANKRD1	
ENSG00000150625	GPM6A	
ENSG00000158578	ALAS2	
ENSG00000166292	TMEM100	
ENSG00000167588	GPD1	
ENSG00000168309	FAM107A	
ENSG00000168484	SFTPC	
ENSG00000169252	ADRB2	
ENSG00000170323	FABP4	
ENSG00000176771	NCKAP5	
ENSG00000179133	C10orf67	
ENSG00000188536	HBA2	
ENSG00000204305	AGER	
ENSG00000206172	HBA1	
ENSG00000206177	HBM	
ENSG00000222761	RNU6-684P	
ENSG00000222979	RN7SKP167	
ENSG00000223587	LINC01986	
ENSG00000224215	LOC105376453	
ENSG00000225427	LINC00331	
ENSG00000226108	RAB28P1	
ENSG00000229604	MTATP8P2	
ENSG00000230352	AL035250.2	
ENSG00000233746	LINC00656	
ENSG00000234646	AL049651.2	
ENSG00000244734	HBB	
ENSG00000259749	NCAPGP2	
ENSG00000260182	AC120498.2	
ENSG00000260940		
ENSG00000261701	HPR	

Lung Genes selected by RF Model		
Gene ID	Gene Name	
ENSG00000183856	IQGAP3	
ENSG00000099953	MMP11	
ENSG00000072163	LIMS2	
ENSG00000168309	FAM107A	
ENSG00000162722	TRIM58	
ENSG00000175874	CREG2	
ENSG00000108370	RGS9	
ENSG00000114859	CLCN2	
ENSG00000144395	CCDC150	
ENSG00000176920	FUT2	
ENSG00000272913	AC009237.14	
ENSG00000114854	TNNC1	
ENSG00000125257	ABCC4	
ENSG00000259918	NDUFA5P11	
ENSG00000253854	AC010834.3	
ENSG00000237163	AC099794.2	
ENSG00000188452	CERKL	
ENSG00000146674	IGFBP3	
ENSG00000269711	AC008763.3	
ENSG00000183019	MCEMP1	
ENSG00000254395	IGHV4-55	
ENSG00000077232	DNAJC10	
ENSG00000254414		
ENSG00000101447	FAM83D	

Lung Genes selected by MI-RF Models		
Gene ID	Gene Name	
ENSG00000183856	IQGAP3	
ENSG00000168309	FAM107A	

Lung Genes selected by RFE-RF Models		
Gene ID	Gene Name	
ENSG00000168309	FAM107A	
ENSG00000114854	TNNC1	

Lung Genes selected by MI-RFE-RF Models		
Gene ID	Gene Name	
ENSG00000168309	FAM107A	

Appendix (3)

Breast Cancer Potential Biomarkers (Genes)

Breast Genes selected by MI-SVC Model			
Gene ID	Gene Name	Gene ID	Gene Name
ENSG00000078596	ITM2A	ENSG00000123975	CKS2
ENSG00000000005	TNMD	ENSG00000235904	RBMS3-AS3
ENSG00000130226	DPP6	ENSG00000163827	LRRC2
ENSG00000072163	LIMS2	ENSG00000279204	AC134043.3
ENSG00000133392	MYH11	ENSG00000134121	CHL1
ENSG00000266524	GDF10	ENSG00000130032	PRRG3
ENSG00000103253	HAGHL	ENSG00000162062	TEDC2
ENSG00000187730	GABRD	ENSG00000132561	MATN2
ENSG00000130176	CNN1	ENSG00000103647	CORO2B
ENSG00000143228	NUF2	ENSG00000088325	TPX2
ENSG00000159433	STARD9	ENSG00000104888	SLC17A7
ENSG00000154065	ANKRD29	ENSG00000167779	IGFBP6
ENSG00000267069	AP005264.1	ENSG00000131471	AOC3
ENSG00000180660	MAB21L1	ENSG00000232079	LINC01697
ENSG00000144642	RBMS3	ENSG00000124440	HIF3A
ENSG00000123612	ACVR1C	ENSG00000133800	LYVE1
ENSG00000120907	ADRA1A	ENSG00000101280	ANGPT4
ENSG00000148180	GSN	ENSG00000187824	TMEM220
ENSG00000198838	RYR3	ENSG00000022267	FHL1
ENSG00000164330	EBF1	ENSG00000162407	PLPP3
ENSG00000152580	IGSF10	ENSG00000168874	ATOH8
ENSG00000137960	GIPC2	ENSG00000277494	GPIHBP1
ENSG00000203805	PLPP4	ENSG00000141338	ABCA8
ENSG00000165124	SVEP1	ENSG00000101938	CHRD1
ENSG00000186466	AQP7P1	ENSG00000169252	ADRB2
ENSG00000263400	TMEM220-AS1	ENSG00000225398	PGM5P4
ENSG00000253250	C8orf88	ENSG00000245812	LINC02202
ENSG00000120049	KCNIP2	ENSG00000099260	PALMD
ENSG00000164087	POC1A	ENSG00000175063	UBE2C
ENSG00000196616	ADH1B	ENSG00000181856	SLC2A4
ENSG00000149124	GLYAT	ENSG00000094963	FMO2
ENSG00000276386	CNTNAP3P2	ENSG00000163815	CLEC3B
ENSG00000170323	FABP4	ENSG00000280339	AP001528.2
ENSG00000128918	ALDH1A2	ENSG00000169418	NPR1
ENSG00000113594	LIFR	ENSG00000148053	NTRK2
ENSG00000081041	CXCL2	ENSG00000119771	KLHL29
ENSG00000158966	CACHD1	ENSG00000177363	LRRN4CL

ENSG00000250961	AC025470.2	ENSG00000078549	ADCYAP1R1
ENSG00000123975	CKS2	ENSG00000241684	ADAMTS9-AS2
ENSG00000235904	RBMS3-AS3	ENSG00000186642	PDE2A
ENSG00000163827	LRRC2	ENSG00000079462	PAFAH1B3
ENSG00000279204	AC134043.3	ENSG00000230838	LINC01614
ENSG00000134121	CHL1	ENSG00000189134	NKAPL
ENSG00000130032	PRRG3	ENSG00000277631	PGM5P3-AS1
ENSG00000162062	TEDC2	ENSG00000101955	SRPX
ENSG00000132561	MATN2	ENSG00000154258	ABCA9
ENSG00000103647	CORO2B	ENSG00000234456	MAGI2-AS3
ENSG00000088325	TPX2	ENSG00000117650	NEK2
ENSG00000104888	SLC17A7	ENSG00000138640	FAM13A
ENSG00000167779	IGFBP6	ENSG00000149451	ADAM33
ENSG00000131471	AOC3	ENSG00000079308	TNS1
ENSG00000232079	LINC01697	ENSG00000165478	HEPACAM
ENSG00000124440	HIF3A	ENSG00000198947	DMD
ENSG00000133800	LYVE1	ENSG00000227467	LINC01537
ENSG00000101280	ANGPT4	ENSG00000170509	HSD17B13
ENSG00000187824	TMEM220	ENSG00000181234	TMEM132C
ENSG00000022267	FHL1	ENSG00000277954	AC092376.2
ENSG00000162407	PLPP3	ENSG00000180914	OXTR
ENSG00000168874	ATOH8	ENSG00000154263	ABCA10
ENSG00000277494	GPIHBP1	ENSG00000105971	CAV2
ENSG00000141338	ABCA8	ENSG00000231246	LINC02884
ENSG00000101938	CHRD1	ENSG00000163431	LMOD1
ENSG00000169252	ADRB2	ENSG00000248869	LINC02511
ENSG00000225398	PGM5P4	ENSG00000266964	FXDY1
ENSG00000245812	LINC02202	ENSG00000177098	SCN4B
ENSG00000099260	PALMD	ENSG00000229246	LINC00377
ENSG00000175063	UBE2C	ENSG00000127564	PKMYT1
ENSG00000181856	SLC2A4	ENSG00000179796	LRRC3B
ENSG00000094963	FMO2	ENSG00000168477	TNXB
ENSG00000163815	CLEC3B	ENSG00000101605	MYOM1
ENSG00000280339	AP001528.2	ENSG00000145777	TSLP
ENSG00000169418	NPR1	ENSG00000091704	CPA1
ENSG00000148053	NTRK2	ENSG00000254862	LGR4-AS1
ENSG00000119771	KLHL29	ENSG00000165810	BTNL9
ENSG00000177363	LRRN4CL	ENSG00000240654	C1QTNF9
ENSG00000078549	ADCYAP1R1	ENSG00000136158	SPRY2
ENSG00000241684	ADAMTS9-AS2	ENSG00000105974	CAV1
ENSG00000186642	PDE2A	ENSG00000149090	PAMR1
ENSG00000079462	PAFAH1B3	ENSG00000196549	MME
ENSG00000230838	LINC01614	ENSG00000165072	MAMDC2
ENSG00000189134	NKAPL	ENSG00000123243	ITIH5

ENSG00000277631	PGM5P3-AS1	ENSG00000154736	ADAMTS5
ENSG00000168497	CAVIN2	ENSG00000123500	COL10A1
ENSG00000137727	ARHGAP20	ENSG00000161649	CD300LG
ENSG00000167434	CA4	ENSG00000168079	SCARA5
ENSG00000099953	MMP11	ENSG00000250692	AC008667.2
ENSG00000165197	VEGFD		

Breast Genes selected by RFE-SVC Model			
Gene ID	Gene Name		
ENSG00000022267	FHL1	ENSG00000250692	AC008667.2
ENSG00000056998	GYG2	ENSG00000154370	TRIM11
ENSG00000108960	MMD		
ENSG00000120907	ADRA1A		
ENSG00000133800	LYVE1		
ENSG00000149124	GLYAT		
ENSG00000154736	ADAMTS5		
ENSG00000156486	KCNS2		
ENSG00000161649	CD300LG		
ENSG00000163710	PCOLCE2		
ENSG00000163792	TCF23		
ENSG00000163815	CLEC3B		
ENSG00000164220	F2RL2		
ENSG00000165197	VEGFD		
ENSG00000166819	PLIN1		
ENSG00000167434	CA4		
ENSG00000168079	SCARA5		
ENSG00000170509	HSD17B13		
ENSG00000174697	LEP		
ENSG00000181104	F2R		
ENSG00000183160	TMEM119		
ENSG00000186205	MTARC1		
ENSG00000186971	KRTAP13-4		
ENSG00000189367	KIAA0408		
ENSG00000231824	AKAIN1		
ENSG00000232079	LINC01697		
ENSG00000240194	CYMP		
ENSG00000253434	LINC02237		
ENSG00000255471	AP001528.1		
ENSG00000259347	AC087482.1		
ENSG00000260102	LINC01070		
ENSG00000272321	AP003355.2		
ENSG00000280339	AP001528.2		

Breast Genes selected by RF Model			
Gene ID	Gene Name	ENSG00000110786	PTPN5
ENSG00000072163	LIMS2	ENSG00000211787	TRAV8-3
ENSG00000259347	AC087482.1	ENSG00000206549	AC109583.1
ENSG00000253250	C8orf88	ENSG00000168101	NUDT16L1
ENSG00000078549	ADCYAP1R1	ENSG00000161905	ALOX15
ENSG00000117724	CENPF	ENSG00000071859	FAM50A
ENSG00000158966	CACHD1	ENSG00000049089	COL9A2
ENSG00000237949	LINC00844	ENSG00000278954	AC130686.1
ENSG00000143228	NUF2	ENSG00000249829	AC016559.2
ENSG00000004776	HSPB6	ENSG00000178623	GPR35
ENSG00000149380	P4HA3	ENSG00000127074	RGS13
ENSG00000243742	RPLP0P2	ENSG00000111704	NANOG
ENSG00000099953	MMP11	ENSG00000207286	Y_RNA
ENSG00000236751	LINC01186	ENSG00000198576	ARC
ENSG00000138639	ARHGAP24	ENSG00000136881	BAAT
ENSG00000119411	BSPRY	ENSG00000241641	RPS23P6
ENSG00000259345	AC013652.1	ENSG00000111224	PARP11
ENSG00000166016	ABTB2	ENSG00000135248	FAM71F1
ENSG00000254862	LGR4-AS1	ENSG00000268455	AC073544.1
ENSG00000178662	CSRNP3	ENSG00000127952	STYXL1
ENSG00000167779	IGFBP6	ENSG00000255191	AP003555.2
ENSG00000171604	CXXC5	ENSG00000198130	HIBCH
ENSG00000259367	AC087286.2	ENSG00000138604	GLCE
ENSG00000186318	BACE1	ENSG00000250461	AC122718.1
ENSG00000109654	TRIM2	ENSG00000250692	AC008667.2
ENSG00000237685	AL139039.3	ENSG00000154370	TRIM11
ENSG00000226527	AP000289.1		
ENSG00000055163	CYFIP2		
ENSG00000105486	LIG1		
ENSG00000213965	NUDT19		
ENSG00000093010	COMT		
ENSG00000250688	TRAJ55		
ENSG00000189134	NKAPL		
ENSG00000160877	NACC1		
ENSG00000196724	ZNF418		
ENSG00000280334	AC009084.2		
ENSG00000229246	LINC00377		
ENSG00000175643	RMI2		
ENSG00000139572	GPR84		

Breast Genes selected by MI-RF Models		
Gene ID	Gene Name	
ENSG00000072163	LIMS2	
ENSG00000253250	C8orf88	
ENSG00000078549	ADCYAP1R1	
ENSG00000158966	CACHD1	
ENSG00000143228	NUF2	
ENSG00000099953	MMP11	
ENSG00000254862	LGR4-AS1	
ENSG00000167779	IGFBP6	
ENSG00000189134	NKAPL	
ENSG00000229246	LINC00377	
ENSG00000250692	AC008667.2	

Breast Genes selected by RFE-RF Models		
Gene ID	Gene Name	
ENSG00000259347	AC087482.1	
ENSG00000072163	LIMS2	
ENSG00000250692	AC008667.2	

Breast Genes selected by MI-RFE-RF Models		
Gene ID	Gene Name	
ENSG00000072163	LIMS2	
ENSG00000250692	AC008667.2	

المخلص

يحدث السرطان عندما تبدأ الخلايا في الانقسام والنمو بشكل كبير لا يمكن السيطرة عليه. السرطان مرض خطير يهدد حياة الإنسان ويؤدي إلى آثار جانبية وخيمة، وفي كثير من الأحيان يؤدي إلى الوفاة. تكمن المشكلة في أن السرطان غالبًا لا يُكتشف مبكرًا في المراحل الأولى، إلا بعد مرور فترة طويلة. وبالتالي، يكون العلاج في هذه الحالة أكثر تعقيدًا وأحيانًا غير ممكن. من هذا المنطلق، فإن الحاجة إلى حل فعال للكشف عن مكان نشوء الورم ونوع السرطان في مرحلة مبكرة أمر ضروري. نظرًا لأن الورم المتنقل، حيث ينتشر الورم من موقعه الأصلي إلى أجزاء مختلفة من الجسم، يصعب علاجه. تتطور الأمراض المختلفة، بما في ذلك السرطان، بسبب الطفرات التي تحدث في التسلسل الجيني للحمض النووي. مكن التقدم التكنولوجي في استخراج تسلسل الجينات للحمض النووي (NGS) علماء الحاسوب وعلماء المعلومات الحيوية من دراسة التسلسل الجيني لتحديد آثار الطفرات الجينية على تكوين أنواع مختلفة من الأمراض. تركز هذه الأطروحة على تنبؤ وتصنيف أنواع مختلفة من السرطانات بناءً على تقنية (NGS) من خلال استخدام بيانات التعبير الجيني (RNA-Seq) باستخدام خوارزميات مختلفة للتعلم الآلي. الهدف الثاني من هذا البحث هو العثور على الجينات التي لها تأثير كبير في تكون أنواع مختلفة من السرطان. تم اقتراح أربعة نماذج مختلفة في هذه الدراسة، وهي (MI-SVM)، (RFE-SVM)، (RF). النموذج الرابع هو نموذج تجميعي جديد يهدف لايجاد الجينات الأكثر صلة بتطور نوع معين من السرطان. في هذا البحث، تم تطبيق أربع قاعدة بيانات واختبارها باستخدام النماذج المطورة، وتم تفسير النتائج والتحقق من صحتها. كما تم إجراء دراسة مقارنة للنتائج التي تم الحصول عليها مع البحث الحالي. تظهر النتائج التجريبية التي وصلت إلى درجات دقة بنسبة 99.4% أن النماذج المقترحة تتنبأ وتصنف السرطان بكفاءة وتكتشف أكثر الجينات المحتملة نسبيًا لتطور نوع معين من السرطان مقارنة بأحدث الدراسات في هذا المجال.

جامعة البحرين

كلية تقنية المعلومات



التنبؤ بالسرطان باستخدام تقنية التسلسل الجيني للحمض النووي وخوارزميات التعلم الآلي

أطروحة مقدمة كجزء من متطلبات الحصول على
درجة الماجستير في تقنية المعلومات

اعداد

مروة عاصم عيفان لطيف

20190175

اشراف

د. رياض بن محمد قسنطيني

استاذ مشارك

د. فاطمة عبدالقادر يوسف البلوشي

استاذ مساعد

جامعة البحرين

مملكة البحرين

مارس 2021